IRE Transactions

ON AUTOMATIC CONTROL

Volume AC-6 FEBRUARY, 1961 Number 1

TABLE OF CONTENTS

PUBLISHED BY THE

PROFESSIONAL GROUP ON AUTOMATIC CONTROL

IRE TRANSACTIONS®
on Automatic Control

George S. Axelby, *Editor*, Air Arm Division,
Westinghouse Electric Corp., Box 746, Baltimore, Md.

Published by The Institute of Radio Engineers, Inc., for the Professional Group on Automatic Control, 1 East 79th Street, New York 21, N. Y. Responsibility for the contents rests upon the authors, and not upon the IRE, the Group or its members. Individual copies of this issue and all available back issues may be purchased at the following prices: IRE members (one copy) $2.25; libraries and colleges $3.25; all others $4.50. Annual subscription price: colleges and public libraries, $12.75; non-members $17.00.

Message from Moscow

IN mid-summer of 1960, from June 27 to July 7, the first International Congress of Automatic Control was held in Moscow at the Moscow State University. There were about 1100 delegates from all over the world, and they came to hear 285 papers concerning control theory, applications, and components as well as to visit various industries and institutes and to meet the Russian scientists and the Russian people.

At the first session of the Congress, we were made aware of the great importance that the Soviet Union is placing on automation and automatic control. The Congress was regarded as a national event. A special postage stamp had been issued to commemorate the event, and Vice Premier A. N. Kosigin, First Vice President of the Council of Ministers of the USSR, welcomed the delegates at the opening session. Later, all delegates and their wives were given engraved invitations to a reception at the Kremlin Palace, a rare event and a special honor.

The principal address at the opening session of the Congress was entitled "Automation and Mankind" and it was presented by Academician V. A. Trapeznikov. It followed welcoming speeches by IFAC President, Professor A. M. Letov of the USSR, by past president Harold Chestnut of the USA, and others. Professor Trapeznikov's talk was widely acclaimed by all the delegates for its sober, yet inspiring survey of the future of automatic control, of the challenging problems that confront us, and of the progress that has been made in Russia and in the Western countries. Professor Trapeznikov was evidently stating the Soviet attitude toward automatic control and the direction it will take when he said, "The goals of automation can only be achieved when all steps of the industrial processes have been automated in all basic industries . . . when comprehensive automation has been realized to cover ramified remote control systems widely using computers . . . not only will automation raise productivity, but it will also radically change the very nature of labor. . . . The full utilization of the benefits arising out of automation, however, is only possible in a rationally organized society where the manpower made redundant due to automation in one field is easily absorbed in others. Our firm conviction—which we do not, of course, impose on anybody—is that this possibility is offered by the socialist system.

". . . With increasingly more material benefits available due to automation, a smaller proportion of people will be needed in production. . . . With more leisure time, . . . man will be able—for the first time in history—to devote to himself the attention he rightly deserves . . . . The goal of automation is a noble one. It is realistic and feasible. But there are a number of obstacles to clear and a number of formidable problems to solve before it can be achieved."

In detail, Professor Trapeznikov outlined the various problems that confront us in the areas of optimal control, adaptive control, and logic machines to supplement human designers. He noted that "advances in engineering and science are laying a single theoretical foundation for the whole range of engineering subjects involved in communication and control. . . . This comprehensive theory—still in the making—is often called either communication and control theory, control theory, or engineering cybernetics. However, the name is not essential. What is important is the crystallization of fundamental ideas, principles, and methods. The development of this single theory is vital to automation.

"No progress of automatic control theory and engineering, however, is possible without commensurate advances in automation hardware. It has always been that every new device . . . has given use to quantitative changes, to further headway of automatic control theory.

"In the field of automation hardware, there are . . . problems which are still waiting for their solution." Professor Trapeznikov emphasized the particular problems of reliability, of obtaining unitized control systems, and microminiaturization or molecular engineering where he said, " . . . The progress made in this field is particularly striking in the Western countries."

Finally, he mentioned the well-known problems of scientific effort and research: " . . . the optimal decision problem as applied to the planning of research and engineering does not and cannot have a hard and fast solution. . . . It is absolutely clear that the role of fundamental research is bound to grow with time, and its scope should be extended. This is the prerequisite for the true progress of science and engineering. On the other hand, the wider the scope . . . the more difficult it is to organize it. How can we possibly exclude cases when the investigator takes a byroad for the highway only to find himself in a blind alley? How can we possibly reduce the drain of effort and time involved in trials and errors in a multitude of unknown domains?

. . . "The prospects of progress before humanity are imposing and advances in automation and control theory are part of the general progress. . . . Peace is the vital prerequisite for the prosperity of man, for the development of science, for the progress of automation."

This talk was a small but significant part of the message that we bring from Moscow. It was more than a learned survey of the control field and its problems—it was a challenge to control engineers to strive toward realization of the promise that is inherent in automatic control. In particular, it was a challenge to the countries of the Western world because it implied that complete industrial automation, and the high standard of living it can create, could not be achieved effectively by any but a socialist form of government. From our direct ob-

servations, it was apparent that the Soviet Union has far to go before it will approach the level of automation that exists in the United States, but it is evident that an accelerated and sincere effort is being made to accomplish their goal of achieving universal automation. In Kiev, for example, a new institute has been built expressly for the purpose of automating all branches of industry. By 1965, it is expected that the institute will have over 6000 engineers working toward this goal, and that there will be several other automatic control institutes in different parts of Russia working intensely, blending theory and practice toward their common goal. If this trend continues successfully, it will be an effort unsurpassed by any nation. With the full support of the government and with the large number of graduate engineers entering the control field each year, the evolution of Russian industry and production could be exceedingly swift.

From the papers and discussions that we heard, from the people that we met, from the industries and institutes that we visited, and from meetings with other delegates, several specific conclusions were reached:

1) The Russians have devoted a greater effort than the United States in developing automatic control theory, and in the future they will contribute far more money and manpower to this field which they regard as a major science.

2) The Russians are more aware of what we are doing in automatic control than we are of what they are doing largely because of their translation methods and their system of distributing this literature. Yet, it appears that their own effort is somewhat hampered by their necessary affiliation with the government and the various overlapping functions of many technological institutes. They have exceptional talent and leadership in the control field, but it appears that it is not yet efficiently organized.

3) The Russians have been leaders for many years in the theory of automatic control, yet it seems that they lag in its application, at least in western Russia. There is also evidence that their progress in developing the theory has been impeded by the fact that it is too mathematical; by not being applied, the practical limitations of the theory are not evident, new problems do not present themselves and the basic engineering need for control becomes lost. However, the Russians appear to realize this, and to instill new life in their control program, they are developing a huge nationwide plan for automating all industries with the leadership of engineers and scientists in new control institutes.

4) To compete with the Russian effort with less manpower and with fewer facilities, we must concentrate on developing basic control theory as derived from actual engineering problems rather than devoting the largest portion of our over-all effort to mathematical manipulation of hypothetical but elegant system equations. Of course, it will be necessary to use and develop new mathematical methods to obtain this theory, but mathematics should be the means and not the end to the development of new engineering knowledge.

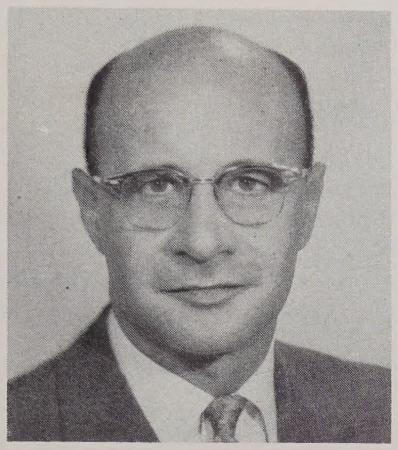
There is more to our message from Moscow:

We were visitors in a land with a language, a culture, and a heritage far different from our own. We met the people there, and, in general, found them to be courteous, helpful, and pleasant. Some delegates thought that there were people who were rather grim and unhappy, but it was not a predominant mood because many of them, especially the younger generation, seemed to be enjoying the new freedoms that have been recently granted to them. In fact, it seemed that they carried the problems of the world more calmly than we do, and they may be less influenced by Soviet proclamations than we are. Of course, they have few of the luxuries that we regard as necessities, but they do have more than they ever had. They have had the promise, now the beginning of hope that their future will be brighter, that one day they may attain and even surpass our standard of living.

Our stay in Russia was limited, our activities were numerous but brief, and our experiences were memorable but veiled with the swift passage of time. Thus, with the specific parts of our message from Moscow there are others less vivid but essential to our over-all impression. Activities, experiences, impressions, memories merge and swirl in a vision: we had glanced behind the iron curtain through a door labeled automatic control. Before us stretched a long corridor paved with mathematical symbols and bordered on either side with high reddish walls, towered and crenelated, dusty and inlaid with glittering relics of a turbulent past. Glimmering torches aligned these ominous walls casting a friendly glow that stretched away in the distance. Far along the way, behind a translucent panel, there was a bustle of activity. We were near enough to see shadowy figures moving behind the panel and to hear the babble of voices and the whirring sound of invisible machinery. In the diffuse light we could not divine its purpose—perhaps it was an illusion—but overhead, above the ancient walls, we could see the dark sky, and from behind the distant panel portentous lights were rising among the stars.

—The Editor

PGAC Chairman's Message*



JOHN M. SALZER†, SENIOR MEMBER, IRE

THE Professional Group on Automatic Control is coming into its own. The ardent work of many an individual both in the national and chapter organizations has borne gratifying, if not remarkable, fruit over the past six years. We have come a long way over a rather short time, as many important achievements can testify.

Presently we count around 4000 members, eleven active chapters, a treasury balance of about $10,000 (in spite of our increased publication costs), we issue at least four TRANSACTIONS each year, and we send the *Newsletter* to all our members several times a year.

The last year was studded with particularly important milestones for PGAC. We participated in the operation of the American Automatic Control Council (AACC), which is a member of the International Federation of Automatic Control (IFAC). The First International Congress was held by IFAC in Moscow during this period. A large number of the 140 American delegates participating in this conference were members of PGAC-IRE. In November, 1959, PGAC sponsored the First National Automatic Control Conference co-sponsored by the other four societies (AIChE, AIEE, ASME, ISA) represented in AACC. Under the Chairmanship of Louis B. Wadel, now our PGAC Vice Chairman, the Dallas Conference proved to be a tremendous success and set the stage for the Annual Joint Automatic Control Conferences. The first of the latter was held in September, 1959, in Cambridge, Massachusetts, under the sponsorship of ASME with appropriate PGAC participation. In the past year, we made two recommendations for IRE awards and one recommendation for a Fellow award. We have also established a new annual award of $100 for the best paper published in the PGAC TRANSACTIONS. For these achievements and for this general progress, PGAC is indebted to all those who helped. The officers and all active members of every chapter, as well as those of us who work on the national committees, had a part in these accomplishments, and we all are thankful for the devoted service of our past Chairman, John E. Ward, who worked with us for two years in this capacity and did so with untiring effort.

We cannot rest on the merits of past accomplishments and on the false assurance that things will take care of themselves. We must continue and strengthen our present activities. We are supporting ISA in the preparation for the next Joint Automatic Control Conference. We continue to be active participants in AACC. We will make proposals for awards and publicize these opportunities to all our members. We will provide more help to the chapters in advice and coordination from the National Chapters Committee, and we expect to establish additional chapters and launch a national membership drive. We will both increase our publications and become more selective in accepting papers. We will continue to sponsor PGAC sessions at the national IRE meetings. We will establish a committee to examine the basic objectives and scope of operation of PGAC and suggest some modification or expansion if appropriate.

These are some of our aims and these are some of the tasks we foresee. All of us will have a part in this program. As your new Chairman, I will try to serve you with diligence and dedication.

* Received by the PGAC, November 17, 1960.

† Director, Intellectronics Labs., Thompson Ramo Wooldridge, Inc., Canoga Park, Calif.

The Issue in Brief

Of the ten papers featured in this issue, half of them are devoted to theoretical optimization of control systems and the rest pertain to application of theory and physical control problems.

***An Optimal Strategy for a Saturating Sampled-Data System,* C. A. Desoer and J. Wing**

A usual second-order sampled-data system is considered with a zero-order hold, but the forcing function applied to the plant may not be larger than unity in absolute value. The problem discussed is that of determining an optimum strategy, a forcing function which forces the plant from an arbitrary initial state of equilibrium in the least number of sampling periods.

***Time Optimal Control of Higher-Order Systems,* F. B. Smith, Jr.**

Phase space has been considered as the basis for synthesizing optimum control for high-order systems, but the mathematical surfaces are difficult to represent physically and this has been considered a handicap. However, this paper considers the determination of an optimum forcing function from the state variables of the system without considering phase space. The theory is developed, an example is given, and a brief discussion follows.

***Control System Performance Measures: Past, Present, and Future,* W. C. Schultz and V. C. Rideout**

This is essentially a review and tutorial paper which explains the development, the use, and the future of expressions which have been proposed to define system performance. The importance of this performance index in the design of adaptive systems is stressed and an extensive bibliography is included.

***A Simulator Study of a Two-Parameter Adaptive System,* R. J. McGrath and V. C. Rideout**

A sinusoidal perturbation is used to determine the optimum performance of a system, and one of the performance indexes discussed in the previous paper gives an indication when the performance deviates from optimum. Two parameters are changed or adapted automatically to changing conditions, and it is shown that the system will adapt itself to an optimum condition when sinusoidal or random inputs are applied to it.

***Optimum Prediction with a Mean Weighted Square Error Criterion,* C. C. Glover**

This paper is an example of the trend for future emphasis on applications of statistical concepts and general integral forms to system optimization that is discussed by W. C. Schultz and V. C. Rideout in another paper in this issue. This theory of optimum prediction is developed and an example is given.

***Control Systems with Minimum Spectral Bandwidth of Plant Input,* J. C. Hung**

It is desirable to attenuate high-frequency components of signals which are applied to the plant of a control system to prevent excitation of inherent resonant poles. It is shown in this paper that minimizing a closed-loop system bandwidth for minimum error does not necessarily minimize the plant input spectrum and the resonant poles may be excited. The conditions which limit the system bandwidth and the plant input spectrum are discussed and an example is present.

***A Network Theory for Carrier-Suppressed Modulated Systems,* G. Weiss**

Considerable literature has been written about carrier-suppressed modulated systems. Some of these previous developments are discussed and unified in this presentation which considers analysis and synthesis of carrier-frequency networks using root-locus techniques and approximate methods. It is shown that approximation methods commonly used are theoretically sound over a relatively wide frequency range.

***Design Aspects of Attitude Control Systems,* M. F. Marx**

This discussion is quite different in nature from those previously described. It concerns the practical problem involved in obtaining a desirable control of missile attitude. It is concluded that the control system must be basically different for various modes and regions of operation. It is suggested that it be adaptive during boost and recovery and that it need not be optimum in the sense of satisfying any particular error criterion.

***Analysis and Design of Feedback System with Gain and Time Constant Variations,* K. Chen**

The problem considered in the paper is encountered in many control system designs. It is desired to determine the transient response of a control loop which has a variable gain and time constant. In this case, the stabilization problem is complicated by the fact that the gain and time constant may become negative and produce open-loop instability. Nevertheless, a method of designing the system to have a desired transient response is discussed in this paper.

***Evaluation of Transient Response Coefficients,* D. S. Billingsley and M. G. Rekoff, Jr.**

A method of obtaining the transient response coefficients from a root-locus plot for roots of any multiplicity is described.

Correspondence

R. M. du Plessis has two contributions: one is a discussion of a previous PGAC paper by Dr. Braun and the other concerns the digital programming of multirate sampled-data systems. P. Dorato discusses short-time stability; and Professor Jury surveys various sampling methods in one contribution and, in another, co-authored by B. H. Bharaucha, he presents some notes on the stability criterion for linear discrete systems.

Information on Translation of Russian Technical Journals

It was noted at the JACC meeting in Boston that complete translations of four Russian journals have been made for the past three years by the ISA and they are available, at low cost, to anyone. Details are given in this issue.

An Optimal Strategy for a Saturating Sampled-Data System*

C. A. DESOER†, SENIOR MEMBER, IRE, AND J. WING†, MEMBER, IRE

Summary—Consider the usual sampled-data control system in which the sampler is followed by a zero-order hold and the transfer function is $G(s)=1/s(s+a)$. Saturation is represented by the fact that the forcing function applied to $G(s)$ may not be larger than 1 in absolute value. The problem is to determine a saturating zero-order hold forcing function which forces the system from an arbitrary initial state to equilibrium in the least number of sampling periods. Such a forcing function is defined as an optimal strategy.

The state plane is divided into boundary states and interior states. To each boundary state corresponds a unique optimal strategy. To each interior state correspond infinitely many optimal strategies.

From the system parameters a polygonal curve, called the critical curve, is defined in the state plane. An optimal strategy is then proposed in which the required forcing function is simply obtained by computing the distance of the representative point in state plane to the critical curve. A simple computer is proposed to implement this optimal strategy. Finally, the proposed optimal strategy is shown to reduce in the limit as $T\to 0$ to that of the corresponding continuous system.

I. Introduction

THE increasing use in control systems of digital data links, digital computers, and other intermittently operative devices has stimulated considerable investigations in the field of sampled-data control systems. One important aspect of the sampled-data field that has been neglected is that problem which corresponds to the optimal relay servo of the continuous control system [1]—[3]. The optimal relay servo problem is concerned with a system described by a linear differential equation with constant coefficients. The control signal of the system $f(t)$ is constrained between the saturation limits $+1$ and -1. The problem is then to determine that $f(t)$ such that the system is forced to equilibrium in minimum time.

The optimal control problem for continuous systems has resulted in considerable effort being expended in its solution as is evident from the large amount of literature available on this subject. This paper extends some preliminary results of Kalman [4]. Here we will be concerned with a particular second-order linear servomechanism. It will be shown in a rigorous manner that the optimal control signal for this system may be obtained by an extremely simple method using standard analog computer techniques [6].

The paper starts by presenting a formulation of the problem that requires the minimum number of parameters. The set of all possible initial states is partitioned into sets R_N, in terms of the minimum number of sampling periods required to reach equilibrium. Theorems 1–3 establish the exact shape of these sets. In Theorem 4, it is shown that only for special initial states is the optimal strategy unique. In Section VI the optimality of the proposed strategy is established, followed by the implementation of the optimal strategy. Finally, the relationship between the optimal strategy for the sampled system and the optimal relay solution for the corresponding continuous system is discussed in detail.

II. Statement of the Problem

Fig. 1 shows the servomechanism that will be considered throughout the paper. Note that the plant is described by a second-order transfer function $G(s)=K/s(s+a)$, $(a>0)$, and that it is preceded by a saturating amplifier. Note also that the feedback loop consists of a computer whose input is $c(t)$ and whose output is $F(t)$. The problem is the following: Assuming that $r(t)$ is zero for all times and given an arbitrary set of initial conditions $c(0)$, $\dot{c}(0)$, find the forcing function $F(t)$ and the corresponding computer which will bring the system to equilibrium in the minimum number of sampling periods.

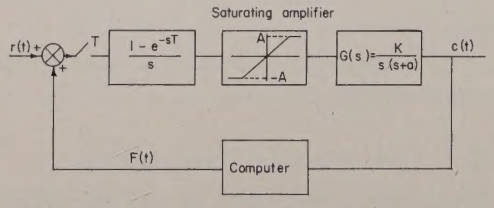


Fig. 1—Block diagram of the system under consideration.

Let us first reduce the number of parameters by appropriate normalizations. By suitably selecting the unit of the forcing function $F(t)$ the saturation limits become $+1$ and -1. This normalization changes the gain constant K of $G(s)$ to K'. By a suitable time normalization the new constant K' of $G(s)$ can be made equal to unity, since by expanding the time scale by a factor of k^{-1} we accomplish two things: 1) change s to sk, and 2) multiply the transfer function by k. This simultaneous forcing function and time normalization reduces the problem to a two-parameter problem: a, the time constant of the plant $G(s)$ and T, the sampling period. Therefore, the problem is reduced to that shown in Fig. 2, where it is understood that the effective forcing function $f(t)$ must satisfy the following two constraints: 1)

* Received by the PGAC, January 18, 1960; revised manuscript received, June 17, 1960. This research was supported by the U. S. Air Force through the Air Force Office of Scientific Research of the Air Res. and Dev. Command, under Contract No. AF 18(1600)-1521.

† Elec. Engrg. Dept., University of California, Berkeley, Calif.

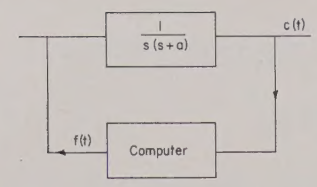


Fig. 2—System under consideration after normalization.

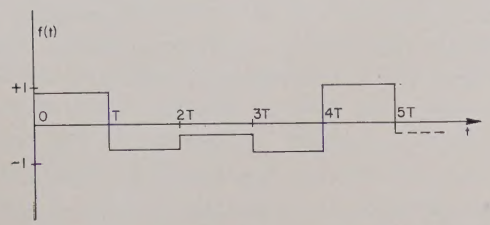


Fig. 3—Typical forcing function $f(t)$.

at all times $|f(t)| \leq 1$, a requirement which follows from the saturation, and 2) $f(t)$ is a zero-order hold function; more precisely, for any interval $kT<t<(k+1)T$, (where k is an integer), the effective forcing function $f(t)$ is constant. This requirement follows from the fact that the original system had a sampler with period T followed by a zero-order hold circuit. Fig. 3 shows a typical forcing function $f(t)$. The effective forcing function is completely defined by the sequence of numbers $f_1, f_2, \cdots$, where f_1 is the value of $f(t)$ during the first sampling period $0<t<T$, and, in general, $f_k=f(t)$ for $(k-1)T<t<kT$, $(k=1, 2, 3, \cdots)$.

The problem can therefore be reformulated as follows: Given the system shown in Fig. 2 and given an arbitary set of initial conditions, $c(0)$ and $\dot{c}(0)$, find an effective forcing function $f(t)$, specified by $f_1, f_2, \cdots$, which will bring the system to equilibrium in the minimum number of sampling periods.

III. Matrix Formulation

Fig. 2 and the constraints on the effective forcing function imply that, for $0<t<T$,

$$\ddot{c}(t) + a\dot{c}(t) = f_1.$$

For initial conditions expressed as $c(0)$ and $\dot{c}(0)$, the solution is

$$c(t) = c(0) + \frac{1 - e^{-at}}{a}\dot{c}(0) + f_1\frac{e^{-at} + at - 1}{a^2} \tag{1}$$

and, consequently,

$$\dot{c}(t) = e^{-at}\dot{c}(0) + f_1\frac{1 - e^{-at}}{a}. \tag{2}$$

Therefore, the relation between the initial conditions for the next sampling period, $c(T)$ and $\dot{c}(T)$, and the present ones is of the form

$$\begin{bmatrix} c(T) \\ \dot{c}(T) \end{bmatrix} = \begin{bmatrix} 1 & \dfrac{1 - e^{-aT}}{a} \\ 0 & e^{-aT} \end{bmatrix}\begin{bmatrix} c(0) \\ \dot{c}(0) \end{bmatrix} + f_1\begin{bmatrix} \dfrac{e^{-aT} + aT - 1}{a^2} \\ \dfrac{1 - e^{-aT}}{a} \end{bmatrix}. \tag{3}$$

Let us denote by $\boldsymbol{A}$ the matrix appearing in (3), namely,

$$\boldsymbol{A} = \begin{bmatrix} 1 & \dfrac{1 - e^{-aT}}{a} \\ 0 & e^{-aT} \end{bmatrix}. \tag{4}$$

Since the relation (3) is fundamental for the rest of this paper, it is worthwhile to change variables in order to give it as simple a form as possible. This change of variable is not indispensable; however, it does simplify considerably the geometric representation of the linear transformation defined by $\boldsymbol{A}$ and consequently helps in visualizing the discussion that follows. For this purpose, let us think in terms of the vector $\mathbf{c}(kT)$ defined by its components $c(kT)$, $\dot{c}(kT)$; where $k=0, 1, 2, \cdots$. The vector $\mathbf{c}(kT)$ can be expressed in terms of the normalized eigenvectors of $\boldsymbol{A}$, $\boldsymbol{e}_1$, $\boldsymbol{e}_2$,

$$\mathbf{c}(kT) = \gamma_1(kT)\boldsymbol{e}_1 + \gamma_2(kT)\boldsymbol{e}_2 \quad (k = 0, 1, 2, \cdots). \tag{5}$$

A standard computation gives

$$\lambda_1 = e^{-aT}, \qquad \boldsymbol{e}_1 = \begin{bmatrix} -\dfrac{1}{\sqrt{1 + a^2}} \\ \dfrac{a}{\sqrt{1 + a^2}} \end{bmatrix};$$

$$\lambda_2 = 1, \qquad \boldsymbol{e}_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Thinking in terms of the vector $\boldsymbol{\gamma}(kT)$ defined by its components $\gamma_1(kT)$, $\gamma_2(kT)$, (3) takes the form

$$\boldsymbol{\gamma}(T) = \boldsymbol{\Lambda}\boldsymbol{\gamma}(0) + f_1\boldsymbol{d} \tag{6}$$

where

$$\boldsymbol{\Lambda} = \begin{pmatrix} e^{-aT} & 0 \\ 0 & 1 \end{pmatrix} \tag{7}$$

and

$$\boldsymbol{d} = \begin{pmatrix} a^{-2}(1 - e^{-aT})(1 + a^2)^{1/2} \\ -Ta^{-1} \end{pmatrix}. \tag{8}$$

It is clear that for the general interval $(k-1)T<t<kT$, (6) reads

$$\boldsymbol{\gamma}(kT) = \boldsymbol{\Lambda}\boldsymbol{\gamma}((k-1)T) + f_k\boldsymbol{d} \quad (k = 1, 2, 3, \cdots). \tag{9}$$

IV. Classification of Initial States

The initial conditions $c(0)$, $\dot{c}(0)$ can also be specified by $\boldsymbol{\gamma}(0)$ as can be readily seen from (5). We shall refer to the vector $\boldsymbol{\gamma}(0)$ as the initial state of the system and think of it as the point $(\gamma_1(0), \gamma_2(0))$ in the (γ_1, γ_2) plane.

It follows from (9) that if the initial state $\boldsymbol{\gamma}(0)$ is such that the system can be brought to equilibrium in N sampling periods, the following relation holds:

$$\boldsymbol{\gamma}(NT) = 0 = \boldsymbol{\Lambda}^N\boldsymbol{\gamma}(0) + f_1\boldsymbol{\Lambda}^{N-1}\boldsymbol{d} + f_2\boldsymbol{\Lambda}^{N-2}\boldsymbol{d} + \cdots + f_{N-1}\boldsymbol{\Lambda}d + f_N\boldsymbol{d} \tag{10}$$

where

$$|f_i| \leq 1 \qquad (i = 1, 2, \cdots, N)$$

and the values of the forcing function $f_1, f_2, \cdots, f_N$ depend on the particular initial state $\boldsymbol{\gamma}(0)$ under consideration.

For $k = 1, 2, 3, \cdots$, define

$$\boldsymbol{r}_k = -\boldsymbol{\Lambda}^{-k}\boldsymbol{d} = \begin{pmatrix} -e^{kaT}a^{-2}(1 - e^{-aT})(1 + a^2)^{1/2} \\ Ta^{-1} \end{pmatrix}. \tag{11}$$

The vectors $\boldsymbol{r}_i$ are shown on Fig. 4. Let us note also that since $a > 0$, the γ_1 components of the $\boldsymbol{r}_i$ increase exponentially with i.

If we now premultiply (10) by $\boldsymbol{\Lambda}^{-N}$ and use (11) we get

$$\boldsymbol{\gamma}(0) = \sum_{i=1}^{N} f_i\boldsymbol{r}, \quad \text{where} \quad |f_i| \leq 1 \text{ for all } i, \tag{12}$$

which is a general representation for the initial states that can be brought into equilibrium in N sampling periods or less. Eq. (12) is the basic relation for the classification of initial states. For convenience, let us introduce a

Definition: R_N' is the set of initial states that can be brought to equilibrium in N sampling periods or less.

We shall presently obtain the shape of the region of R_N'. For convenience, we shall detail the steps of the reasoning in the form of theorems.

Theorem 1: The region R_N' is convex, *i.e.*, if two initial states represented by the points P_1 and P_2 can be brought to equilibrium in N sampling periods or less, the same is true for any initial state on the line segment P_1P_2.

Proof: Let $\boldsymbol{OP}_1$ be represented by $\boldsymbol{\gamma}(0)$. Since P_1 lies in R_N',

$$\boldsymbol{\gamma}(0) = \sum_{i=1}^{N} f_i\boldsymbol{r}_i \text{ where } |f_i| \leq 1 \ (i = 1, 2, \cdots, N). \tag{13}$$

Similarly, let $\boldsymbol{OP}_2$ be represented by $\boldsymbol{\gamma}'(0)$. Then since P_2 lies in R_N',

$$\boldsymbol{\gamma}'(0) = \sum_{i=1}^{N} f_i'\boldsymbol{r}_i \text{ where } |f_i'| \leq 1 \ (i = 1, 2, \cdots, N). \tag{14}$$

Let us write $f_i' = f_i + \Delta f_i$; consequently, any point on the line segment P_1P_2 can be represented by

$$\sum_{i=1}^{N} (f_i + \delta\Delta f_i)\boldsymbol{r}_i \quad \text{where} \quad 0 \leq \delta \leq 1. \tag{15}$$

The inequalities in (13) and (14) imply that for any δ in the interval $0 \leq \delta \leq 1$

$$|f_i + \delta\Delta f_i| \leq 1 \qquad (i = 1, 2, \cdots, N);$$

hence, the vector sum (15) has the standard form (12) irrespective of the value of δ, and represents a point in R_N'.

Q.E.D.

We can now precisely define the region R_N' by its boundary. This is done by

Theorem 2: R_N' is the closed set whose boundary is the convex polygon Π_N which has the following $2N$ vertices:[1]

$$\begin{aligned}
\boldsymbol{OP}_1 &= \boldsymbol{r}_1 - \boldsymbol{r}_2 - \boldsymbol{r}_3 - \cdots - \boldsymbol{r}_N \\
\boldsymbol{OP}_2 &= \boldsymbol{r}_1 + \boldsymbol{r}_2 - \boldsymbol{r}_3 - \cdots - \boldsymbol{r}_N \\
&\vdots \\
\boldsymbol{OP}_N &= \boldsymbol{r}_1 + \boldsymbol{r}_2 + \boldsymbol{r}_3 + \cdots + \boldsymbol{r}_N \\
\boldsymbol{OP}_{-1} &= -\boldsymbol{r}_1 + \boldsymbol{r}_2 + \boldsymbol{r}_3 + \cdots + \boldsymbol{r}_N \\
\boldsymbol{OP}_{-2} &= -\boldsymbol{r}_1 - \boldsymbol{r}_2 + \boldsymbol{r}_3 + \cdots + \boldsymbol{r}_N \\
&\vdots \\
\boldsymbol{OP}_{-N} &= -\boldsymbol{r}_1 - \boldsymbol{r}_2 - \boldsymbol{r}_3 - \cdots - \boldsymbol{r}_N
\end{aligned} \tag{16}$$

where O is the origin of the (γ_1, γ_2) plane.

[1] Figs. 6–8 illustrate the case $N+2$, 3 and 4.

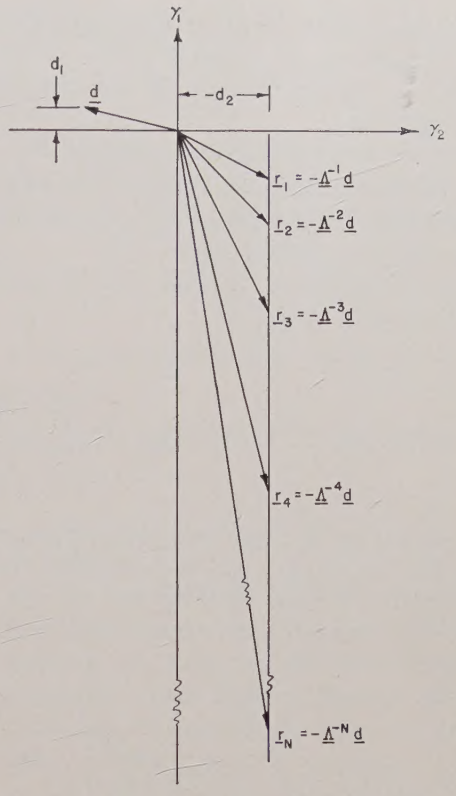


Fig. 4—Illustration of the vectors $\boldsymbol{r}_i$ and $\boldsymbol{d}$.

Proof: Let us observe that P_k is the point symmetric to P_{-k} with respect to the origin O of the plane (γ_1, γ_2). Therefore, the convex polygon Π_N has the origin as a center of symmetry.

Let P be an arbitrary interior point of R_N' (see Fig. 5). The line drawn from the origin O through P intersects the edges P_kP_{k+1} and $P_{-k}P_{-(k+1)}$ of Π_N at the points Q_k and Q_{-k}, respectively.

From the definition of P_k, P_{k+1}, P_{-k}, $P_{-(k+1)}$ we have the following representations:

$$\mathbf{OQ}_k = \mathbf{r}_1 + \mathbf{r}_2 + \cdots + \mathbf{r}_k + \delta\mathbf{r}_{k+1} - \mathbf{r}_{k+2} - \cdots - \mathbf{r}_N \qquad -1 \leq \delta \leq 1 \quad (17)$$

$$\mathbf{OQ}_{-k} = -\mathbf{r}_1 - \mathbf{r}_2 - \cdots - \mathbf{r}_k + \delta\mathbf{r}_{k+1} + \mathbf{r}_{k+2} + \cdots + \mathbf{r}_N \qquad -1 \leq \delta \leq 1. \quad (18)$$

Upon examination of these two expressions with (12), it follows that Q_k and Q_{-k} can be brought to equilibrium in N sampling periods. Since P lies on the segment Q_kQ_{-k}, by Theorem 1 the same holds true for P.

This establishes the fact that any interior point of Π_N and any boundary point of Π_N belongs to R_N'. In short, any point in Π_N belongs to R_N'. It remains to establish the converse property, namely, that no point outside the polygon Π_N can belong to R_N'.

Consider now an arbitrary point P outside Π_N. We show that P cannot be brought to the origin in N sampling periods or less, *i.e.*, it cannot be expressed in the form of (12) and hence does not belong to R_N'. Suppose P could be brought to the origin in exactly N sampling periods and no less. Then

$$\mathbf{OP} = f_1''\mathbf{r}_1 + f_2''\mathbf{r}_2 + \cdots + f_k''\mathbf{r}_k + f_{k+1}''\mathbf{r}_{k+1} + f_{k+2}''\mathbf{r}_{k+2} + \cdots + f_N''\mathbf{r}_N \quad (19)$$

where

$$|f_i''| \leq 1.$$

Since P is outside Π_N, OP intersects the boundary of Π_N at some point Q. Let Q be on the edge P_kP_{k+1}. Then

$$\mathbf{OQ} = \mathbf{r}_1 + \mathbf{r}_2 + \cdots + \mathbf{r}_k + \delta\mathbf{r}_{k+1} - \mathbf{r}_{k+2} - \cdots - \mathbf{r}_N$$

where

$$-1 \leq \delta \leq 1.$$

Clearly, one representation for the f_i'''s of (19) is

$$f_i'' = +\frac{|\mathbf{OQ}|}{|\mathbf{OP}|} > 1 \qquad i = (1, 2, \cdots, k)$$

$$f_{k+1}'' = \delta\frac{|\mathbf{OQ}|}{|\mathbf{OP}|} \qquad -1 \leq \delta \leq 1$$

$$f_i'' = -\frac{|\mathbf{OQ}|}{|\mathbf{OP}|} < -1 \quad i = (k+2, \cdots, N).$$

Thus, all f_i'''s with the possible exception of f''_{k+1} have absolute values greater than unity. It will be possible to bring P to equilibrium in N sampling periods (as assumed above) only if there exists some modification of the f_i'', say $f_i''+\Delta_i$, such that if the f_i'''s are replaced by the $f_i''+\Delta_i$'s (19) still holds, and $|f_i''+\Delta_i| \leq 1$ for $i=1, 2, 3, \cdots, N$. Hence, by subtraction,

$$\sum_{i=1}^{k} \Delta_i\mathbf{r}_i + \Delta_{k+1}\mathbf{r}_{k+1} + \sum_{k=2}^{N} \Delta_i\mathbf{r}_i = 0$$

with

$$\Delta_i < 0 \quad \text{for} \quad i = 1, 2, \cdots, k$$

$$\Delta_i > 0 \quad \text{for} \quad i = k+2, k+3, \cdots, N.$$

Hence,

$$\Delta_{k+1}\mathbf{r}_{k+1} = \sum_{i=1}^{k} |\Delta_i|\,\mathbf{r}_i - \sum_{k+2}^{N} |\Delta_i|\,\mathbf{r}_i.$$

This relationship cannot hold irrespective of the sign of Δ_{k+1}. Observe that the angles of the vectors $\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N$ (see Fig. 4) with the axis γ_2 increase monotonically with the subscript i of $\mathbf{r}_i$. Thus, if $\mathbf{p}_{k+1}$ is a unit vector perpendicular to $\mathbf{r}_{k+1}$ and such that

$$\mathbf{p}_{k+1}\mathbf{r}_i > 0 \quad \text{for} \quad i = 1, 2, \cdots, k,$$

then,

$$\mathbf{p}_{k+1}\mathbf{r}_i < 0 \quad \text{for} \quad i = k+2, k+3, \cdots, N.$$

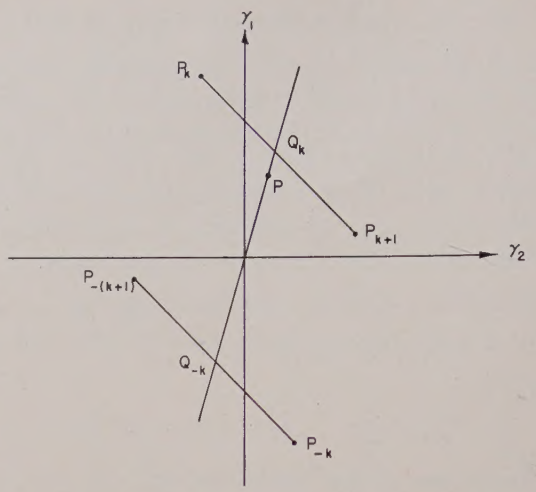


Fig. 5—Illustration for the proof of Theorem 2.

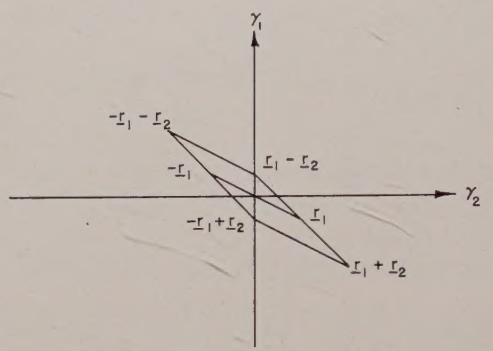


Fig. 6—The region R_1' and R_2' with the vertices shown explicitly.

Therefore, if we take the scalar product of both sides of the equality by $\boldsymbol{p}_{k+1}$, the left-hand side is zero and the right-hand side is >0. This contradiction proves that assumption that P could be brought to equilibrium in N sampling periods and be outside Π_N is false.

Q.E.D.

The polygon Π_{N+1} has $2(N+1)$ vertices P_1', P_2', $\cdots$, P'_{N+1}, P'_{-1}, P'_{-2}, $\cdots$, $P'_{-(N+1)}$. From (16) it follows that a) for $k=1, 2, \cdots, N$, $\boldsymbol{OP}_k'$ is obtained from $\boldsymbol{OP}_k$ by adding $-\boldsymbol{r}_{N+1}$ and $\boldsymbol{OP}'_{-k}$ by adding $\boldsymbol{r}_{N+1}$ to $\boldsymbol{OP}_{-k}$; b) $\boldsymbol{OP}'_{N+1}=\boldsymbol{OP}_N+\boldsymbol{r}_{N+1}$ and $\boldsymbol{OP}'_{-(N+1)}=\boldsymbol{OP}_{-N}-\boldsymbol{r}_{N+1}$. Since the regions R_N' and R'_{N+1} are defined completely by the polygon Π_N and Π_{N+1}, respectively, we have a recursion rule for obtaining R'_{N+1} from R_N':

The boundary of R'_{N+1} is obtained by adding $\pm\boldsymbol{r}_{N+1}$ in an outward direction to the boundary of R_N'.

A stronger statement that is implied in Theorem 2 is that any point P in R'_{N+1} but not in R_N' can be written as

$$\boldsymbol{OP} = \boldsymbol{OQ} + f_{N+1}\boldsymbol{r}_{N+1} \quad \text{with} \quad |f_{N+1}| \leq 1 \qquad (20)$$

where Q is a point on the boundary of Π_N, and $f_{N+1}\boldsymbol{r}_{N+1}$ points in the outward direction. For example, in Fig. 7, it is apparent that the boundary of R_3' is obtained from that of R_2' by adding $\boldsymbol{r}_3$ in an outward direction, and in Fig. 8, that of R_4' is obtained from that of R_3' by adding $\boldsymbol{r}_4$ in an outward direction. For convenience we wish to introduce the

Definition: R_N is the set of all initial states that can be brought to the origin in N sampling periods and *no less*. From the definition of R_N and R_N' we have the obvious

Theorem 3: $R_N=R_N'-R'_{N-1}$, *i.e.*, R_N is obtained by deleting from R_N' all the points that belong to R'_{N-1}.

V. Optimal Strategies

We define an *optimal strategy* by the following requirement: given a $\boldsymbol{\gamma}(0)$ that is a point of R_N, an optimal strategy is *any* effective forcing function $f(t)$, specified by $f_1, f_2, \cdots, f_N$, that brings the point $\boldsymbol{\gamma}(0)$ to the origin in exactly N sampling periods.

Let us observe that since the two vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are linearly independent it follows that if either $\boldsymbol{\gamma}(0)$ lies in R_1 or R_2, the optimal strategy is *unique*. This is a consequence of the fact that the two linear algebraic equations in f_1 and f_2 implied by

$$\boldsymbol{\gamma}(0) = f_1\boldsymbol{r}_1 + f_2\boldsymbol{r}_2$$

have a unique solution.

We note also that if $\boldsymbol{\gamma}(0)$ belongs to R_N with $N\geq 3$, the corresponding set of two linear algebraic equations implied by

$$\boldsymbol{\gamma}(0) = f_1\boldsymbol{r}_1 + f_2\boldsymbol{r}_2 + \cdots + f_N\boldsymbol{r}_N$$

may have more than one solution. In order to proceed further, we define

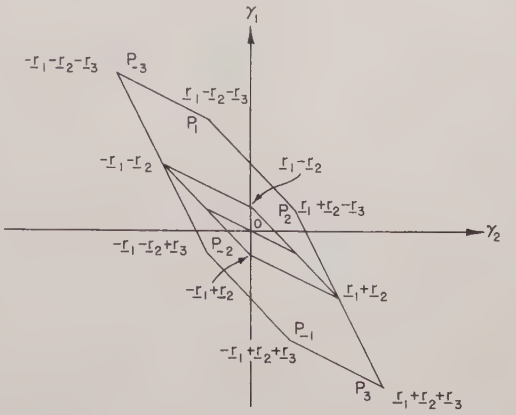


Fig. 7—The region R_1', R_2' and R_3' with the vertices shown explicitly.

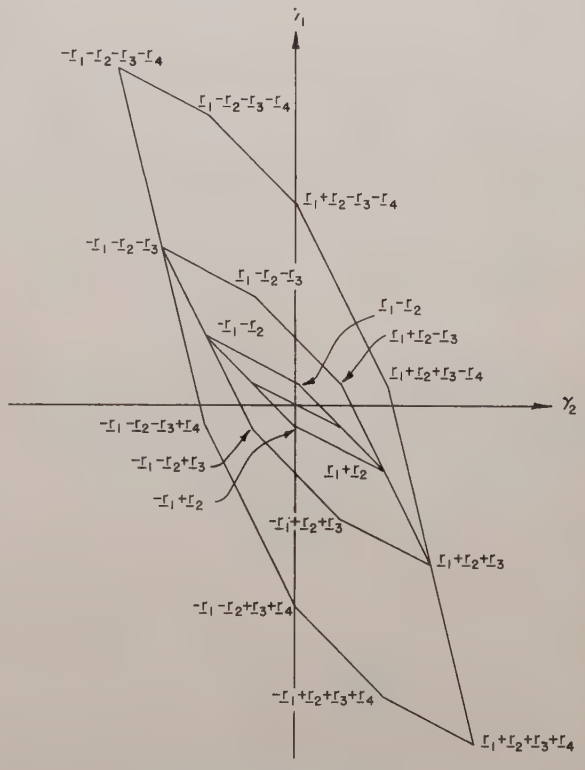


Fig. 8—The region R_1', R_2', R_3' and R_4' with the vertices shown explicitly.

The outer boundary of R_N consists of the edges $P_{-N}P_1$, P_1P_2, $\cdots$, $P_{N-1}P_N$ and the edges $P_{+N}P_{-1}$, $P_{-1}P_{-2}$, $\cdots$, $P_{-(N-1)}P_{-N}$.

An interior state of R_N is any state not on the outer boundary of R_N.

A boundary state of R_N is any state on the outer boundary of R_N.

Theorem 4: Let N be an integer larger than 2. For any $\boldsymbol{\gamma}(0)$ that is an interior state of R_N, there exists an infinite number of optimal strategies. For any $\boldsymbol{\gamma}(0)$ that is a boundary state of R_N, there is a unique optimal strategy.

To prove Theorem 4 we rewrite (6)

$$\gamma(0) = \Lambda^{-1}\gamma(T) - f_1\Lambda^{-1}d,$$

and using (11)

$$\gamma(0) - f_1 r_1 = \Lambda^{-1}\gamma(T). \tag{21}$$

If $\gamma(0)$ is in R_N, then by the definition of the optimal strategy, an optimal value of f_1 is such that $\gamma(T)$ is in R_{N-1}. An equivalent form of (21) expressed in terms of regions R_N and R_{N-1} is

$$R_N - f_1 r_1 = \Lambda^{-1}R_{N-1}. \tag{22}$$

The interpretation of (22) is as follows: Start with R_{N-1}, determine $\Lambda^{-1}R_{N-1}$; R_N is then generated by translating $\Lambda^{-1}R_{N-1}$ by all possible $f_1 r_1$, with $|f_1| \leq 1$. Conversely, if $\gamma(0)$ is in R_N, there is an f_1, where $|f_1| \leq 1$, such that $\gamma(0) - f_1 r_1$ is in $\Lambda^{-1}R_{N-1}$. An optimal value of the forcing function during the first sampling interval is then any such value of f_1.

In order to determine $\Lambda^{-1}R_{N-1}$, we have from Theorem 3 that

$$R_{N-1} = R_{N-1}' - R_{N-2}'$$

and therefore,

$$\Lambda^{-1}R_{N-1} = \Lambda^{-1}R_{N-1}' - \Lambda^{-1}R_{N-2}'. \tag{23}$$

Since R'_{N-1} and R'_{N-2} are convex, $\Lambda^{-1}R'_{N-1}$ and $\Lambda^{-1}R'_{N-2}$ are convex; hence, we need only consider the vertices of $\Lambda^{-1}R'_{N-1}$ and $\Lambda^{-1}R'_{N-2}$. In evaluating (23) we will take a specific value of N, say $N=4$. The regions R_2', R_3' and R_4' are shown in Figs. 7–9. The cross-hatched area in Fig. 9 is $\Lambda^{-1}R_3$. Table I gives a complete tabulation of the vertices. In the table use was made of the fact that

$$\Lambda^{-1}r_n = r_{n+1} \qquad (n = 1, 2, \cdots)$$

which is a consequence of (11).

In Fig. 10 we take an arbitrary interior initial state $\gamma(0)$, which is in R_4, and show how it is translated by $f_1 r_1$ with $-1 \leq f_1 \leq 1$. Clearly, f' is an optimal value of the forcing function since $\gamma(0) - f' r_1$ is in $\Lambda^{-1}R_3$. Similarly, f'' is an optimal value of the forcing function since $\gamma(0) - f'' r_1$ is in $\Lambda^{-1}R_3$. It is apparent that any f_1 in the range $f' \leq f_1 \leq f''$ is an optimal value of the forcing function. If $\gamma(0)$ is an interior state of R_4, $f' \neq f''$ and, hence,

TABLE I

Vertices of Convex Region R_4'	
$r_1 - r_2 - r_3 - r_4$	$-r_1 + r_2 + r_3 + r_4$
$r_1 + r_2 - r_3 - r_4$	$-r_1 - r_2 + r_3 + r_4$
$r_1 + r_2 + r_3 - r_4$	$-r_1 - r_2 - r_3 + r_4$
$r_1 + r_2 + r_3 + r_4$	$-r_1 - r_2 - r_3 - r_4$
Vertices of Convex Region R_3'	
$r_1 - r_2 - r_3$	$-r_1 + r_2 + r_3$
$r_1 + r_2 - r_3$	$-r_1 - r_2 + r_3$
$r_1 + r_2 + r_3$	$-r_1 - r_2 - r_3$
Vertices of Convex Region R_2'	
$r_1 - r_2$	$-r_1 + r_2$
$r_1 + r_2$	$-r_1 - r_2$
Vertices of Convex Region $\Lambda^{-1}R_3'$	
$r_2 - r_3 - r_4$	$-r_2 + r_3 + r_4$
$r_2 + r_3 - r_4$	$-r_2 - r_3 + r_4$
$r_2 + r_3 + r_4$	$-r_2 - r_3 - r_4$
Vertices of Convex Region $\Lambda^{-1}R_2'$	
$r_2 - r_3$	$-r_2 + r_3$
$r_2 + r_3$	$-r_2 - r_3$

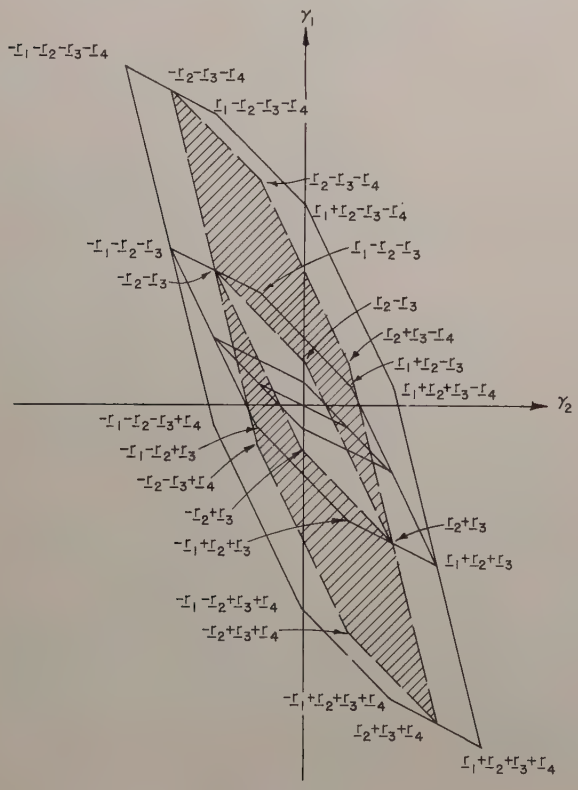


Fig. 9—The cross-hatched area is the set $\Lambda^{-1}R_3$.

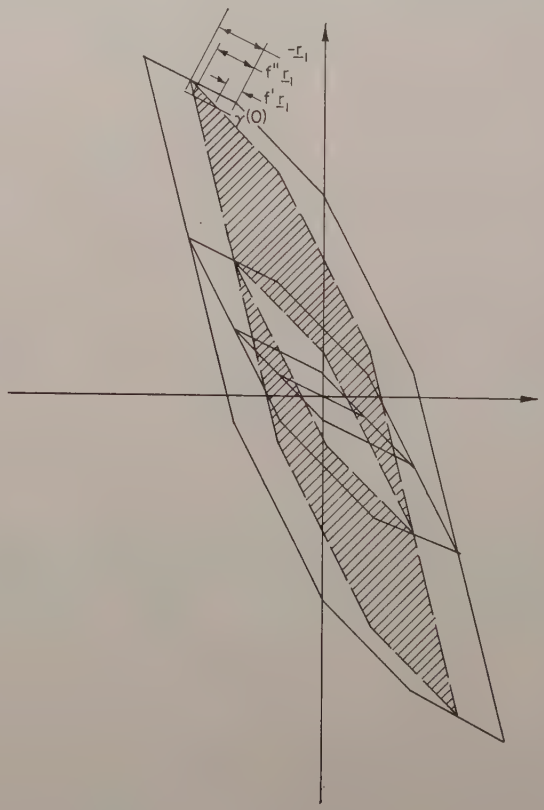


Fig. 10—The definition of f', f'' in terms of $\gamma(0)$ and of the set $\Lambda^{-1}R_3$.

there is an infinite number of optimal values for f_1. If Q is on the outer boundary of R_N then, either

Case 1: it is on an edge of the form

$$P_kP_{k+1} \quad \text{or} \quad P_{-k}P_{-(k+1)} \qquad (k = 1, 2, \cdots, N-1)$$

or

Case 2: it is on edge $P_{-N}P_1$ or P_NP_{-1}.

In the first case,

$$\boldsymbol{OQ} = \boldsymbol{r}_1 + \boldsymbol{r}_2 + \cdots + \boldsymbol{r}_k + \delta\boldsymbol{r}_{k+1} - \boldsymbol{r}_{k+2} - \cdots - \boldsymbol{r}_N$$

where $-1 \leq \delta \leq 1$ and $1 \leq k \leq N-1$ (or the same expression with the sign reversed). From this representation of $\boldsymbol{OQ}$, it is clear that $f_1 = +1$ is an optimal strategy. Furthermore, for such a state Q, $-1 < f_1 < 1$ cannot be an optimal strategy: If $|f_1| < 1$ were applied, at the next sampling period the point Q would be in Q_1 with, as a consequence of (6),

$$\boldsymbol{OQ}_1 = \boldsymbol{\Lambda}\boldsymbol{OQ} + f_1\boldsymbol{d}.$$

From (11),

$$\boldsymbol{OQ}_1 = -(1 - f_1)\boldsymbol{d} + \boldsymbol{r}_1 + \boldsymbol{r}_2 + \cdots + \boldsymbol{r}_{k-1} + \delta\boldsymbol{r}_k - \boldsymbol{r}_{k+1} - \cdots - \boldsymbol{r}_{N-1}.$$

Define Q_2 by

$$\boldsymbol{OQ}_2 = \boldsymbol{r}_1 + \boldsymbol{r}_2 + \cdots + \boldsymbol{r}_{k-1} + \delta\boldsymbol{r}_k - \boldsymbol{r}_{k+1} - \cdots - \boldsymbol{r}_{N-1};$$

thus,

$$\boldsymbol{OQ}_1 = \boldsymbol{OQ}_2 + (1 - f_1)(-\boldsymbol{d}).$$

Q_2 is on the edge $P_{k-1}P_k$ of Π_{N-1}. Since $1 - f_1 > 0$, and since $-\boldsymbol{d}$ makes a smaller angle with the γ_2 axis than the edge $P_{k-1}P_k$, Q_1 is outside Π_{N-1}. Hence, $f_1 < 1$ cannot be an optimal policy, and $f_1 = +1$ is the unique optimal policy.

In case 2, to be specific, assume Q to be on the edge $P_{-N}P_1$. From the representation (16) of $\boldsymbol{OP}_{-N}$ and $\boldsymbol{OP}_1$,

$$\boldsymbol{OQ} = \delta\boldsymbol{r}_1 - \boldsymbol{r}_2 - \boldsymbol{r}_3 - \cdots - \boldsymbol{r}_N \quad \text{where} \quad -1 \leq \delta \leq 1.$$

Clearly, δ is an optimal strategy. Furthermore, this optimal strategy is unique, *i.e.*, if $f_1 \neq \delta$, f_1 cannot be an optimal strategy. This can be shown by following exactly the same procedure as the one used in case 1.

Q.E.D.

VI. The Proposed Optimal Strategy

The object of this section is to establish an optimal strategy which is easy to instrument and which is valid for any state belonging to R_N, where N is an arbitrary positive integer.

Let us start by putting emphasis on two particular polygonal curves:

The critical curve is obtained by joining successively the vertices defined by

$$\cdots, -\sum_{i=2}^{N} \boldsymbol{r}_i, \cdots, -\boldsymbol{r}_2, +\boldsymbol{r}_2, \cdots, \sum_{i=2}^{N} \boldsymbol{r}_i, \cdots$$

(see Fig. 11);

The polygonal curve K is obtained by joining successively the vertices defined by

$$\cdots, -\sum_{i=1}^{N} \boldsymbol{r}_i, \cdots,$$

$$-\boldsymbol{r}_2 - \boldsymbol{r}_1, -\boldsymbol{r}_1, +\boldsymbol{r}_1, \boldsymbol{r}_1 + \boldsymbol{r}_2, \cdots, \sum_{i=1}^{N} \boldsymbol{r}_i, \cdots.$$

Theorem 5: If $\boldsymbol{\gamma}(0)$ is in R_N it can be expressed either as

$$\boldsymbol{\gamma}(0) = \boldsymbol{r}_1 + \boldsymbol{r}_2 + \cdots + \boldsymbol{r}_k + \delta\boldsymbol{r}_{k+1} - \boldsymbol{r}_{k+2} - \cdots - \boldsymbol{r}_{N-1} - \delta_1\boldsymbol{r}_N \qquad (24)$$

or as

$$\boldsymbol{\gamma}(0) = -\boldsymbol{r}_1 - \boldsymbol{r}_2 - \cdots - \boldsymbol{r}_k + \delta\boldsymbol{r}_{k+1} + \boldsymbol{r}_{k+2} + \cdots + \boldsymbol{r}_{N-1} + \delta_1\boldsymbol{r}_N \qquad (25)$$

where $-1 \leq \delta \leq 1$, $0 < \delta_1 \leq 1$, $0 \leq k \leq N-1$. Note that the representation (24) holds for all points to the right of the polygonal curve K and (25) holds for those that lie to the left of the polygonal curve K.

Proof: Observe first that from (17) and (18) any boundary state of R_{N-1} can be written either as

$$\boldsymbol{r}_1 + \boldsymbol{r}_2 + \cdots + \boldsymbol{r}_k + \delta\boldsymbol{r}_{k+1} - \boldsymbol{r}_{k+2} - \cdots - \boldsymbol{r}_{N-1} \qquad (17')$$

or as

$$-\boldsymbol{r}_1 - \boldsymbol{r}_2 - \cdots - \boldsymbol{r}_k + \delta\boldsymbol{r}_{k+1} + \boldsymbol{r}_{k+2} + \cdots + \boldsymbol{r}_{N-1} \qquad (18')$$

where $-1 \leq \delta \leq 1$ and $0 \leq k \leq N-1$. Observe further that $\boldsymbol{\gamma}(0)$ being in R_N implies that the point $\boldsymbol{\gamma}(0)$ can be obtained from a boundary point of R_{N-1} by adding in an

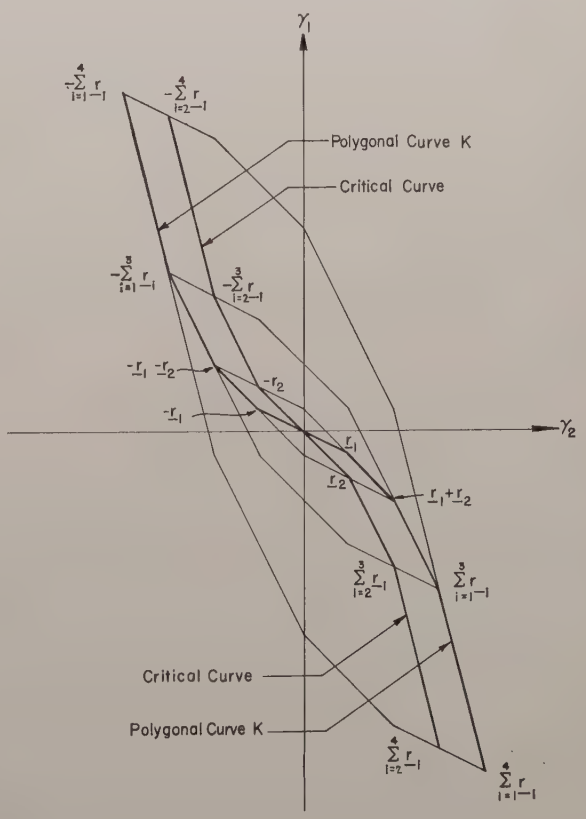


Fig. 11—The initial curve and the polygonal curve K.

outward direction the vector $\delta_1 r_N$ with $|\delta_1| \leq 1$. [See (20).] The relations (24) and (25) simply reflect that particular fact. For the boundary states of R_{N-1} represented by (17′) [(18′), respectively] the outward direction is given by $-\delta_1 r_N$ ($+\delta_1 r_N$, respectively) with $0 < \delta_1 \leq 1$.

Q.E.D.

The proposed optimal strategy follows directly from the canonical representations (24), (25). To get started consider those initial states for which $k=0$ in (24) and (25), *i.e.*,

$$\gamma(0) = \delta r_1 - r_2 - r_3 - \cdots - r_{N-1} - \delta_1 r_N \qquad -1 \leq \delta \leq 1$$

$$\gamma(0) = \delta r_1 + r_2 + r_3 + \cdots + r_{N-1} + \delta_1 r_N$$

or

$$\gamma(0) - \delta r_1 = -r_2 - r_3 - \cdots - r_{N-1} - \delta_1 r_N \qquad -1 \leq \delta \leq 1$$

$$\gamma(0) - \delta r_1 = r_2 + r_3 + \cdots + r_{N-1} + \delta_1 r_N. \tag{26}$$

In both equations (26), the right-hand side represents a point on the critical curve. Observe that

1) If $k>0$, the optimal strategy implied by (24), (25) requires $f_1 = \pm 1$;
2) If $k=0$ and if $|\delta| = 1$, then the optimal strategy implied by (24) and (25) requires $f_1 = \delta$, *i.e.*, $|f_1| = 1$;
3) If $k=0$ and if $|\delta| < 1$, then $f_1 = \delta$ where δ is such that $\gamma(0) - \delta r_1$ be a point on the critical curve.

This leads to the following rule for determining f_1.

Rule: Compute δ' such that $\gamma(0) - \delta' r_1$ be a point on the critical curve. If $\delta' \geq 1$, take $f_1 = 1$, where f_1 is the effective forcing function for the first sampling period. If $\delta' \leq -1$, take $f_1 = -1$. If $-1 < \delta' < 1$, take $f_1 = \delta'$.

This rule is the only rule required. This follows from Bellman's principle of optimality [5] since at each sampling instant one may take the point of view that it is a new problem that is just starting; so that at each sampling instant the problem is to determine the optimal f_1 which is precisely what the rule above accomplishes.

VII. Implementation of the Proposed Optimal Strategy

The proposed optimal strategy is completely defined by the rule of the preceding section: Thus, at each sampling instant we need to compute δ' such that $\gamma(0) - \delta' r_1$ be a point on the critical curve.

The first step is to transform the critical curve from the (γ_1, γ_2) plane to the $(c, \dot{c})$ plane, since $c(t)$ is the input to the computer. From Section III it follows that the vectors $s_1, s_2, \cdots, s_N, \cdots$, of the $(c, \dot{c})$ plane corresponding to the vectors $r_1, r_2, \cdots, r_N, \cdots$, of the (γ_1, γ_2) plane are given by

$$s_k = E r_k \qquad (k = 1, 2, 3, \cdots) \tag{27}$$

where E is the matrix which has the eigenvectors e_1 and e_2 as columns. These vectors are shown in Fig. 12. Similarly, the critical curve can be drawn in the $(c, \dot{c})$ plane. However, since the proposed optimal strategy requires the determination of δ' such that $c(0) - \delta' s_1$ be a point of the critical curve, it is more convenient to rotate the coordinates and use the axes OX_1 and OX_2, shown in Fig. 12, which are orthogonal, where OX_1 is the support of s_1. Thus, in the (x_1, x_2) coordinates, the determination of δ' will amount to taking a difference of abscissas.

The transformation law between the vector $c(0)$ of the $(c, \dot{c})$ plane to the vector $x(0)$ of the (x_1, x_2) plane is the rotation

$$x(0) = T c(0)$$

where

$$T = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Similarly, the vectors s_k' of the (x_1, x_2) plane which corresponds to the vectors s_k of the $(c, \dot{c})$ plane are obtained by

$$s_k' = T s_k \qquad (k = 1, 2, \cdots).$$

The critical curve (*i.e.*, the curve that joins successively the vertices defined by

$$\cdots - \sum_2^N s_i', \cdots, -s_3' - s_2', -s_2', s_2', s_2' + s_3', \cdots, \sum_2^N s_i', \cdots)$$

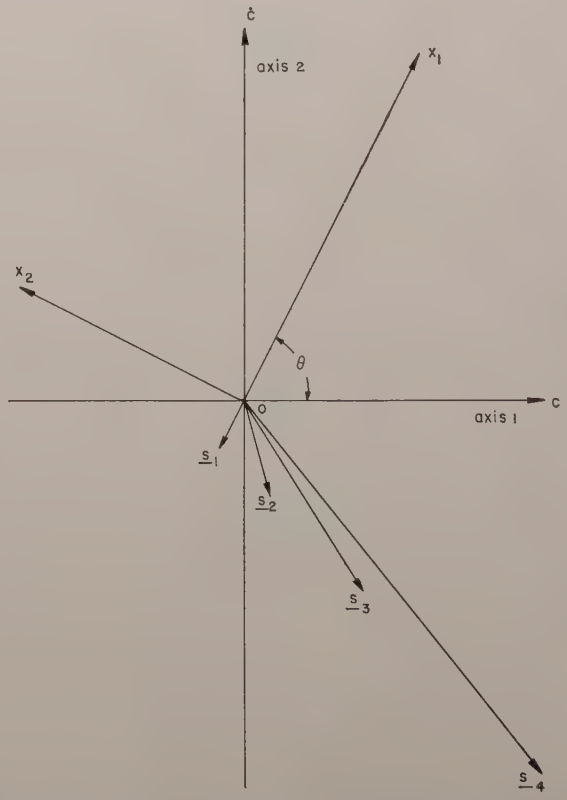


Fig. 12—The new axes Ox_1, Ox_2, Ox_3 shown with respect to the old axes.

in the (x_1, x_2) plane is shown in Fig. 13. Also illustrated is the determination of δ' for a given $\mathbf{x}(0)$: If $f(x_2(0))$ is the abscissa of the critical curve corresponding to the ordinate $x_2(0)$ of $\mathbf{x}(0)$, then

$$\delta' = \frac{f(x_2(0)) - x_1(0)}{|\mathbf{s}_1'|}.$$

Fig. 14 shows a block diagram of the computer: 1) from $c(t)$, $\dot{c}(t)$ is obtained by differentiation; 2) using $c(0)$ and $\dot{c}(0)$, $\mathbf{x}(0)$ is obtained by performing the linear transformation $\mathbf{x}(0) = \boldsymbol{T}\mathbf{c}(0)$ (see detailed analog computer diagram, Fig. 15); 3) $f(x_2(0))$, the abscissa of the critical curve is obtained by the function generator, shown in detail in Fig. 16; 4) from δ' the saturating amplifier gives f_1, the optimal effective forcing function.

The details of the function generator are shown in Fig. 16 [6]. The input to the function generator is $x_2(0)$ and $-x_2(0)$. The output of the function generator is $-f(x_2)$ which is the critical curve reflected about the x_2 axis. This is the dashed curve shown in Fig. 17. The output of Amplifier A_7 is then the required value of δ'.

The method by which the reflected critical curve is broken up in order that it may be generated by diode circuits as summation of monotonic functions is shown in Fig. 17. Each of the monotonic functions M_i is specified by a breakpoint BP_i and a slope as specified by the angle ϕ_i. The relationship of BP_i's and ϕ_i's with the critical curve is clearly shown in Fig. 17. For each M_i to be generated a diode circuit with its associated breakpoint and the slope control is required. Fig. 16 shows the standard physical setup.

VIII. Comparison with the Continuous Case

It is well known [1]–[3] that the optimal strategy for a second-order continuous system consists of applying maximum positive or maximum negative excitation at all times till equilibrium is reached. The change from one to the other excitation takes place at a definite time. The precise instant at which switching takes place is that instant when the representative phase plane point reaches the switching curve. This curve consists of that trajectory which terminates at the origin when maximum positive excitation is applied and that trajectory which terminates at the origin when maximum negative excitation is applied. The strategy for the continuous system is then to apply that maximum excitation to bring the state of the system to the switching curve and, at that instant, reverse the sign of the excitation.

In the following we are going to prove a theorem exhibiting the relation between the optimal strategy for the sampled system and that known one for the corresponding continuous system. As is expected intuitively, the proposed optimal strategy for the sampled system coincides in the limit as $T \to 0$ with that of the continuous system.

Before stating precisely the relation between the optimal strategies let us recall the polygonal curve:

The polygonal curve K is obtained by joining successively the vertices defined by

$$\cdots - \sum_1^N \mathbf{r}_i, \cdots, -\mathbf{r}_2 - \mathbf{r}_1, -\mathbf{r}_1, \mathbf{r}_1, \mathbf{r}_1 + \mathbf{r}_2, \cdots, + \sum_1^N \mathbf{r}_i, \cdots.$$

The vertices of the polygonal curve K are shown on Fig. 11.

Reference to Theorem 5 and (24) shows that, typically, the optimal forcing function is $+1$ for a number of sampling periods, then δ with $|\delta| \leq 1$, and then -1 for the remaining sampling periods except the last where it is $-\delta_1$. If we would relabel 0 the instant at the end of the sampling period during which the forcing function was δ, one would have

$$\boldsymbol{\gamma}(0) = \mathbf{r}_1 + \mathbf{r}_2 + \cdots + \mathbf{r}_{N-k-1} + \delta_1 \mathbf{r}_{N-k};$$

that is, by following the optimal strategy, the state point ends up by approaching the origin by jumping along the polygonal curve K. With this in mind, we state

Theorem 6: Consider the sampled system shown in Fig. 1, together with the corresponding continuous system. Let S be the switching curve associated with the continuous system. Then 1) for every value of T, the sampling period of the sampled system, the vertices of the polygonal curve K lie on the switching curve S. Thus, for every T, the polygonal curve K is a piecewise linear approximation to the switching curve S. 2) As $T \to 0$, the proposed optimal strategy for the sampled system becomes identical to that of the continuous system.

Proof: A general vertex point from the bottom half of the polygonal curve K of Fig. 11 is given by

$$\boldsymbol{OV}_N = \sum_{n=1}^N \mathbf{r}_n = \sum_{n=1}^N \begin{bmatrix} \dfrac{-e^{anT}\sqrt{a^2+1}\,(1-e^{-aT})}{a^2} \\ T/a \end{bmatrix}.$$

Therefore, if the initial state is V_N, it will require N sampling periods or NT seconds to reach the origin. Suppose now that the sampling period T is reduced to zero, while all the other parameters of the system remain unchanged. In order to keep the initial state fixed, we must make $N \to \infty$ as $T \to 0$ in such a manner that NT remains equal to its initial value t_0. In the limit we have

$$\lim_{T\to 0} \sum_{n=1}^{N=t_0/T} \mathbf{r}_n = \lim_{T\to 0} \sum_{n=1}^{N=t_0/T} \begin{bmatrix} -\dfrac{T\sqrt{a^2+1}\,e^{anT}}{a} \\ T/a \end{bmatrix}.$$

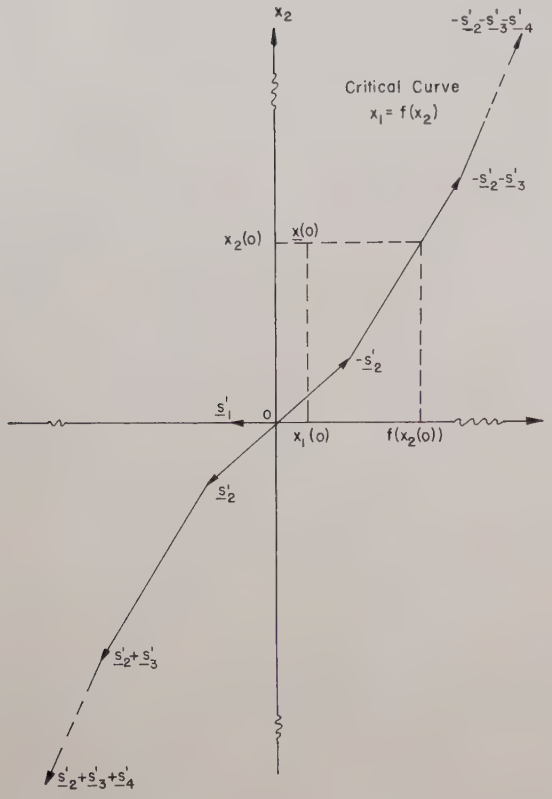


Fig. 13—The critical curve $x_1=f(x_2)$ in the Ox_1x_2 plane in terms of the $\mathbf{s}_i$''s.

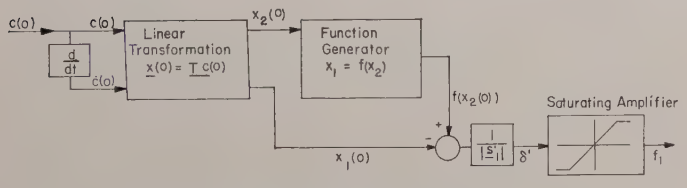


Fig. 14—Block diagram of the computer generating the optimal strategy.

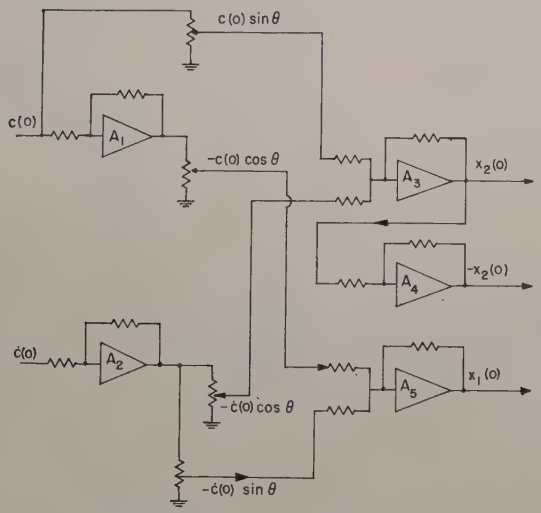


Fig. 15—Detailed analog computer block diagram for the transformation $\mathbf{x}=T\mathbf{c}$.

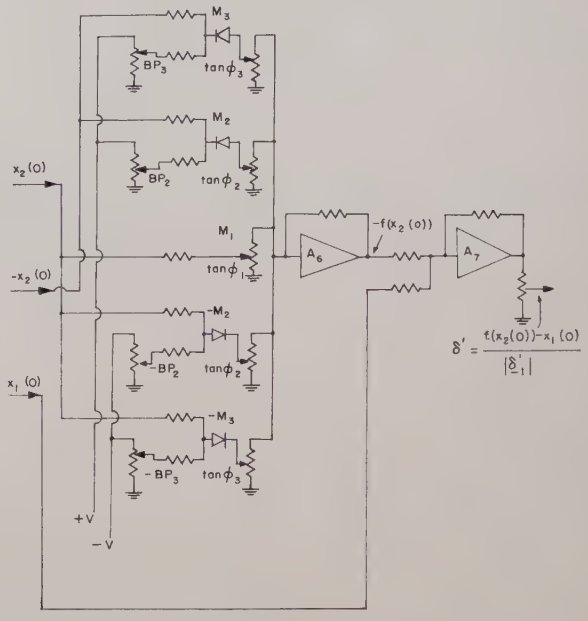


Fig. 16—Function generator simulating the critical curve $x_1=f(x_2)$.

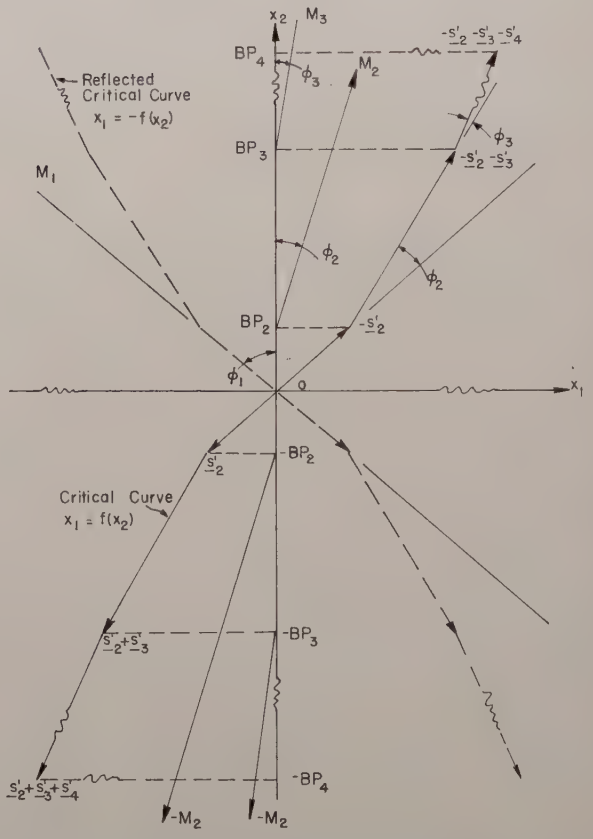


Fig. 17—Illustration of the method for generating the reflected critical curve as a sum of monotonic functions.

The right-hand side is recognized to be in the form of a Riemann sum; hence, as $T \to 0$, it is equal, in the limit, to the integral

$$\int_0^{t_0} \begin{bmatrix} -\dfrac{\sqrt{a^2+1}\, e^{ax}}{a} \\ 1/a \end{bmatrix} dx.$$

The lower limit of integration is zero since

$$\lim_{T \to 0} r_1 = 0.$$

Performing the integration we get

$$\begin{bmatrix} \dfrac{\sqrt{a^2+1}\,(1 - e^{at_0})}{a^2} \\ t_0/a \end{bmatrix}. \tag{28}$$

Let us now show that, provided $NT = t_0$, the point V_N remains fixed in the plane as $T \to 0$

$$OV_N \equiv \sum_{n=1}^{N} r_N$$

$$= \begin{bmatrix} \dfrac{\sqrt{a^2+1}\,(e^{-aT} - 1)}{a^2} (e^{aT} + e^{2aT} + \cdots + e^{NaT}) \\ NT/a \end{bmatrix}$$

or

$$OV_N = \begin{bmatrix} \dfrac{\sqrt{a^2+1}\,(1 - e^{NaT})}{a^2} \\ NT/a \end{bmatrix}. \tag{29}$$

This expression is equal to the value of the integral (28) evaluated at $t_0 = NT$. Therefore, for any T and N, the vertex V_N of the polygonal curve K lies on the continuous curve specified by (28).

We now claim that (28) represents the switching curve of the continuous system in the (γ_1, γ_2) coordinates. To establish this fact, the following procedure will be followed: (29) is transformed to the $(c, \dot{c})$ coordinates by the linear transformation $\boldsymbol{E}$ of (27). The result is

$$c = \frac{e^{at_0} - at_0 - 1}{a^2}$$

$$\dot{c} = \frac{1 - e^{at_0}}{a}.$$

By eliminating t_0 one obtains the equation

$$c = -\frac{\dot{c}}{a} - \frac{1}{a^2} \ln (1 - a\dot{c}) \tag{30}$$

of a curve of the $(c, \dot{c})$ plane which, irrespective of the value of T, goes through the vertices of the polygonal curve K.

It remains to show that the curve (30) is the switching curve of the corresponding continuous system. This is readily done by integrating the differential equation

$$\frac{d\dot{c}}{dt} + a\dot{c} = 1$$

and requiring that the trajectory pass through the origin. The result is

$$c = -\frac{\dot{c}}{a} - \frac{1}{a^2} \ln (1 - a\dot{c})$$

which is identical to (30). Thus, the first assertion of the theorem is established.

As $T \to 0$, $r_1 \to 0$, therefore the critical curve and the polygonal curve K become identical in the limit to the switching curve S. Therefore, in the limit as $T \to 0$, the proposed optimal strategy requires that the forcing function be $+1$ for states to the left of the switching curve S and -1 for states to the right of the switching curve S, as can be seen from Fig. 17. In other words, as $T \to 0$, the proposed optimal strategy for the sampled system becomes identical to that of the continuous system.

Q.E.D.

IX. Conclusion

This paper has solved the problem of finding, proving and implementing the optimal strategy for the sampled-data servo of Fig. 1. The key to the solution was the general representation (24), (25) of the states from which the origin can be reached in N sampling periods and no less. The implementation of the proposed optimal strategy is shown in Fig. 14. It is interesting to note that in the second-order sampled data case, the forcing function is not always as large as possible in absolute value. In the continuous case, it can be shown [7]–[9] under very general conditions that the optimal forcing function always has its absolute value as large as possible.

Bibliography

[1] D. C. McDonald, "Nonlinear techniques for improving servo performance," *Proc. Natl. Electronics Conf.*, vol. 6, pp. 400–421; 1950.

[2] A. M. Hopkin, "A phase-plane approach to the design of saturating servo-mechanism," *Trans. AIEE*, vol. 70, pt. I, pp. 631–639; 1951.

[3] D. Bushaw, "Optimal discontinuous forcing terms," in "Contributions to the Theory of Nonlinear Oscillations," Princeton University Press, Princeton, N. J., vol. 4, pp. 29–52; 1958.

[4] R. E. Kalman, "Optimal nonlinear control of saturating systems by intermittent action," 1957 IRE Wescon Convention Record, pt. 4, pp. 130–135.

[5] R. E. Bellman, "Dynamic Programming," Princeton University Press, Princeton, N. J.; 1957. Also in "Modern Mathematics for the Engineer," E. F. Beckenbeck, Ed., McGraw-Hill Book Co., Inc., New York, N. Y.; 1956.

[6] G. A. Korn and T. M. Korn, "Electronic Analog Computers," McGraw-Hill Book Co., Inc., New York, N. Y. ch. 6; 1956.

[7] C. A. Desoer, "The bang bang servo problem treated by variational techniques," *Information and Control*, vol. 2, 4, pp. 333–349; December, 1959.

[8] R. E. Bellman, I. Glicksberg and O. Gross, "On the bang bang control problem," *Quart Appl. Math.*, vol. 14, pp. 11–18; January, 1956.

[9] J. P. LaSalle, "Time optimal control systems," *Proc. Natl. Acad. Science* (USA), vol. 45, pp. 573–577; April, 1959.

Time-Optimal Control of Higher-Order Systems*

FRED B. SMITH, JR.†

Summary—**Practical extension of time-optimal control to systems of higher order than three has been limited primarily by difficulties in physically representing surfaces in a phase space of these higher dimensions. A method is presented here for obtaining the forcing function as a function of the state variables without requiring use of the phase space concept. On line solution of a set of transcendental equations is required. Results of a digital simulation of a fourth-order, real-root, single-degree-of-freedom system are presented. In a digital solution the system operates as a series of short open-loop control intervals. The effect of including derivatives of the input for prediction is shown for second-order model inputs.**

Introduction

DURING the last ten years there has been considerable interest shown in the literature in improving the characteristics of basically linear systems by introducing a nonlinearity into the control loop. A common nonlinearity introduced has been a relay representing a saturation or maximum permitted forcing function. When used with a linear switching function, it has led to a practical high-gain system exhibiting many of the characteristics of an adaptive system [1]. A second characteristic which has received attention is that of response time, subject to a limited forcing function. Under various titles—"Optimum Relay Control System," "Bang-Bang Control Problems," "Predictor Switching"—the concern has been that of bringing the error and its $n-1$ derivatives to zero in minimum time after a step displacement subject to a limited forcing function. The early work of Hopkin [2], McDonald [3], and Bushaw [4] in the phase plane rather thoroughly investigated second-order systems and showed their feasability. The works of LaSalle [5], Bellman, Glicksberg, and Gross [6], and Bass [7] generalized this and put time-optimal, or minimum response time regulation on a firm mathematical foundation for systems of arbitrary order subject to certain conditions on the roots. Recently, under the instigation of Pontriagin, a number of Russian investigators have approached the problem from a different aspect and have added an interesting geometrical interpretation of the optimum control problem [8]–[11]. While these works added considerable insight into the problem and showed the existence and nature of the solutions, they have not as yet yielded a simple method for obtaining practical control systems of high order.

There are at least two difficultities encountered when attempting to obtain higher-order time optimal regulation or control. The first is that of obtaining the forcing function sign as a function of the state variables. The usual approach has been through the phase space concept [13]. In principle this is a straight-forward extension of the phase plane analysis which is so useful in second-order nonlinear problems. In practice, however, it is not so simple. As the order of the system increases, the mathematical manipulation required to eliminate time from the solutions becomes complex, and the hardware required to express these switching surfaces becomes prohibitive.

Although a number of third-order systems have been done using electro-optical, two-variable function generators, two single-variable function generators properly switched, and digital computers for obtaining the phase space switching surfaces [12]–[16], none of these seem practical to extend beyond third order. The recent method of Kurzweil [16] may be a possible exception to this.

The second difficulty which has been mentioned as being a block in the way of extension to higher orders is that of measurement of the controlled variables. The usual transfer function approach to single-variable control synthesis leads one to talk about the controlled variable and its $n-1$ derivatives as making up the state vector to be controlled. Because of noise, it is not usually possible to measure directly by differentiation more than one or possibly two derivatives. However, high-order systems are generally made up of coupled first- or second-order systems. It is these physical variables which appear in the equations of motion as initially derived, and they can be measured. If it is desired to control a single variable and its $(n-1)$ derivatives instead of controlling the variables independently, then it is usually possible to write the derivatives as a linear combination of these physical variables and the forcing function [17], [18]. This paper does not treat the interesting problem of what to control, but concentrates on a method for determining the forcing function.

A Nonphase Space Method for Obtaining Optimum Switching

Hopkin and Wang [20], before going on to solve their problem in phase space, mentioned briefly the possibility of using switching times for determining the forcing function. Recently, Lee [19] suggested in more detail a technique for obtaining the forcing function sign without requiring the concept of switching surfaces in the phase space. The method presented here is essentially his, but utilizes some of the properties of time-optimal switching as proved by him and others to simplify the numerical computation required. Briefly, the technique is as follows.

* Received by the PGAC, June 6, 1960; revised manuscript received, October 17, 1960. This paper was presented at the 7th Region IRE Conf., Seattle, Wash., May 25, 1960.

† Minneapolis-Honeywell Regulator Co., Minneapolis, Minn.

Consider the system to be controlled as described by the set of first-order differential equations:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1n} \\ \vdots \\ a_{n1} \cdots a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} u. \qquad (1)$$

Here the x_i's are the variables to be controlled—either the physical variables of the system or linear combinations of them. Only a single forcing function u is considered and is restricted to be $|u| = 1$ for all time less than the response time. In more compact notation, (1) is written

$$\dot{x} = Ax + Bu. \qquad (1a)$$

Now to simplify integration of these equations, we transform to principal coordinates, *i.e.*, find a transformation matrix S such that

$$S^{-1}AS = D, \text{ a diagonal matrix.}$$

Making the substitution

$$V = S^{-1}x, \qquad \beta = S^{-1}B$$

in (1a), we obtain the principal coordinate equations of motion,

$$V = DV + \beta u. \qquad (2)$$

The solution of these equations is

$$V(T) = \Phi(T)V(o) + \int_0^T \Phi(T)\Phi^{-1}(\tau)\beta u d\tau, \qquad (3)$$

where $\Phi(T) = e^{DT}$ is the fundamental matrix of the system $\Phi^{-1}(\tau) = \Phi(-\tau) = e^{-D\tau}$ for this constant coefficient system. From previous studies it is known that for time-optimal control, the forcing function u will take on values of only $u = \pm 1$ for time less than T, and for the case of real distinct roots will change sign at most $(n-1)$ times during the time interval zero to T. What is not known, and indeed what is the heart of the problem, is whether its initial value should be *plus* one or *minus* one. To solve this we consider both cases, assuming that the forcing function will be made zero at time T. Multiplying (3) by $\Phi^{-1}(T)$, and breaking the integral into n sections with constant forcing functions changing sign in successive sections, one obtains

$$\Phi^{-1}(T)V(T) - V(o) = \pm u\left[\int_0^{t_1} \beta\Phi^{-1}(\tau)d\tau - \int_{t_1}^{t_2} \beta\Phi^{-1}(\tau)d\tau \cdots (-1)^{n-1}\int_{t_{n-1}}^{T} \Phi^{-1}(\tau)\beta d\tau\right]. \qquad (4)$$

Since

$$\Phi^{-1}(\tau) = e^{-D\tau} = \begin{bmatrix} e^{-\alpha_1\tau} & & & 0 \\ & e^{-\alpha_2\tau} & & \\ & & \ddots & \\ 0 & & & e^{-\alpha_n\tau} \end{bmatrix},$$

(4) when integrated becomes

$$\begin{aligned}
1 + \frac{\pm(-\alpha_1)}{\beta_1 u}\left[e^{-\alpha_1 T}V_1(T) - V_1(o)\right] &= 2e^{-\alpha_1 t_1} - 2e^{-\alpha_1 t_2} \cdots (-1)^{n-1}e^{-\alpha_1 T}, \\
1 + \frac{\pm(-\alpha_2)}{\beta_2 u}\left[e^{-\alpha_2 T}V_2(T) - V_2(o)\right] &= 2e^{-\alpha_2 t_1} - 2e^{-\alpha_2 t_2} \cdots (-1)^{n-1}e^{-\alpha_2 T}, \\
&\vdots \\
1 + \frac{\pm(-\alpha_n)}{\beta_2 u}\left[e^{-\alpha_n T}V_n(T) - V_n(o)\right] &= 2e^{-\alpha_n t_1} - 2e^{-\alpha_n t_2} \cdots (-1)^{n-1}e^{-\alpha_n T}.
\end{aligned} \qquad (5)$$

The n unknowns, $t_1, \cdots t_{n-1}, T$ are the times at which the forcing function must change sign so that at time T the system will be at the desired $V(T)$, starting from $V(o)$ at time $= 0$. If the n equations assuming an initial plus u are solved for a set of positive ordered switching tines, *i.e.*, $0 < t_1 < t_2 \cdots t_{n-1} < T$, and the same is done for an initial negative sign, then the sign corresponding to the set of equations with the minimum T is the correct optimum initial sign for that set of initial conditions. Since for a system with real distinct roots and at most $(n-1)$ switches, the switching surfaces to arrive at the origin are unique [13], the sets of equations (7) can have only one set of positive-ordered roots.

If now the system could be assumed to have a constant $V(T)$, and were subject to no external disturbance, then the system could be operated open-loop with the forcing function changing sign at the times $t_1, t_2, \cdots t_{n-1}$ and being turned off at T. This assumption is not usually the case with practical systems. Consequently, it is necessary to consider each point in time (time now) as time zero, and to resolve the equations with the present value of coordinates as initial conditions obtaining a new choice sign. This is illustrated in Fig. 1 for a system with pure inertia. It is assumed that $x(T) = 0$, $x_2(o) = 0$, $x_1(o) = 0.7$; *i.e.*, we want to return to the origin after a step displacement of x_1. Curve A shows the path of the system if there are no external disturbances. Curve B shows the path when the system is subjected to a constant external force of 0.5 for a period of 0.8 second. Fig. 2 shows the computed switching times for these two conditions as a function of time.

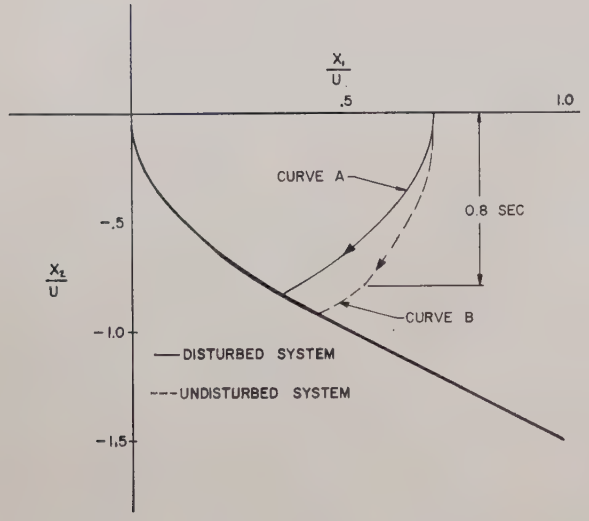


Fig. 1—Phase trajectories of pure inertia-undisturbed and disturbed by constant external forcing function for 0.8 second.

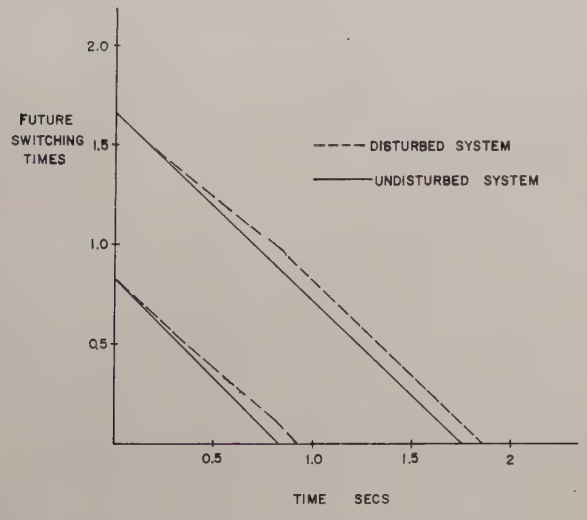


Fig. 2—Predicted switching times along disturbed and undisturbed trajectories.

Simulation of a Fourth-Order System

As an example of the way in which this method is used, consider the fourth-order single-degree-of-freedom system:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\\ \dot{x}_4\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0.413 & 2.051 & -0.268\end{bmatrix} \cdot \begin{bmatrix}x_1\\ x_2\\ x_3\\ x_4\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 0\\ 1\end{bmatrix} u. \quad (6)$$

It has a transfer function of the form

$$\frac{k}{s(s+a_1)(s+a_2)(s+a_3)},$$

which might represent positional control of a motor through a filtered power amplifier. Roots of the system were chosen to be $+1.40$, -1.47, -0.2. The coordinates being controlled are x_1 and its three derivatives. The implicit switching equations (5) for this system are

$$\pm\frac{V_1(o)}{u} = 2t_1 - 2t_2 + 2t_3 - T \pm \frac{V_1(T)}{u},$$

$$\left[1 \pm \frac{0.2V_2(o)}{u}\right] = 2e^{0.2t_1} - 2e^{0.2t_2} + 2e^{0.2t_3} - \left[1 \pm \frac{0.2V_2(T)}{u}\right] e^{0.2T},$$

$$\left[1 \pm \frac{1.47V_3(o)}{u}\right] = 2e^{1.47t_1} - 2e^{1.47t_2} + 2e^{1.47t_3} - \left[1 \pm \frac{1.47V_3(T)}{u}\right] e^{1.47T},$$

$$\left[1 \mp \frac{1.40V_4(o)}{u}\right] = 2e^{-1.40t_1} - 2e^{-1.40t_2} + 2e^{-1.40t_3} - \left[1 \pm \frac{1.40V_4(T)}{u}\right] e^{-1.40T}. \quad (7)$$

Most investigators in the past have put $V(T)=0$, *i.e.*, have worked the regulation problem. Hopkin and Wang [20] in their second-order investigation assumed the input $V(t)$ could be adequately described by a three-term expansion with the coefficients being the present measured value of V and its first two derivatives. This will be done here, and the effect of including these prediction terms will be shown for one type of input.

Although it is felt that (7) can be solved in analog fashion, the present simulations have been done digitally, with (7) being solved by a simple iteration procedure. The exponentials are approximated by a two-term series expansion (keeping linear terms in the switching times), and for each iteration the switching times are changed by an increment proportional to the increment which would make the linear approximation equal to zero. No convergence problems were experienced with this real root system. The only difficulty occurs when the initial guess for the switching times is considerably less than the positive-ordered set which satisfies the equations. Convergence then occurs on an unordered solution. Because the linear approximation is singular when two switching times are equal (which occurs on the switching boundary), it is necessary to work slightly off the exact switching surface. This results in a small steady-state limit cycle.

Because of the finite time required to obtain an iterative solution, there is a delay from the time the coordinates are read in until a corresponding solution is found. During this time the system is operating open-loop on the previous solution. In truth then, the sampled system operates as a series of short open-loop

intervals. This differs from the usual sampled-data systems where the forcing term changes only at the sampling time. It results in good control as long as external forces do not change switching times much during the open-loop intervals. Of course, the smaller the expected outside influences, the slower can be the computer. For most of the traces presented here, the IBM-650 was scaled to one millisecond per iteration. For the inputs shown, an average of 5–6 iterations was required for each new set of coordinates. For inputs of Figs. 6 to 8, switching times were all maintained at less than one second.

Fig. 3 is a rough schematic of the closed-loop system.

Fig. 4 is a plot of switching times vs step input of x_1/u. Fig. 5 shows the time response to a step input of $x_1/u = 0.01$. The effect of having to switch slightly later than the correct time because of singularities on the switching surface mentioned earlier, is seen in the x_3 and x_4 traces. This is the result of only 0.5-msec lag in the true switching time. A lag of 1 msec requires about twice the number of switches to obtain equivalent errors in x_3 and x_4. Effect on transient response is negligible.

Fig. 6 shows response of this fourth-order time-optimal regulator (solid line) to a step input through a second-order model with 0.5 damping and 1.5 rad/sec natural frequency (dashed line). No derivatives of the input have been included in the implicit switching equa-

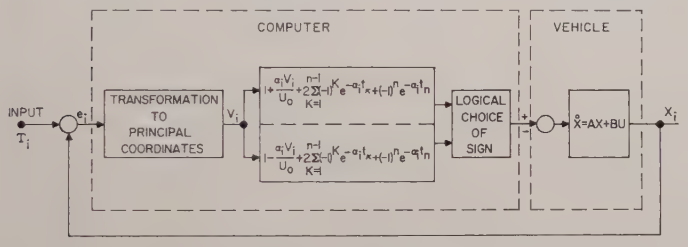


Fig. 3—General time-optimal control loop.

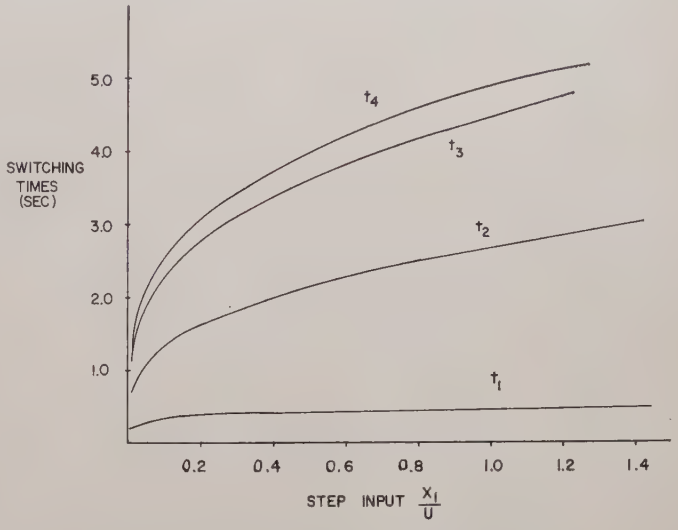


Fig. 4—Switching times for step input of x_1/u.

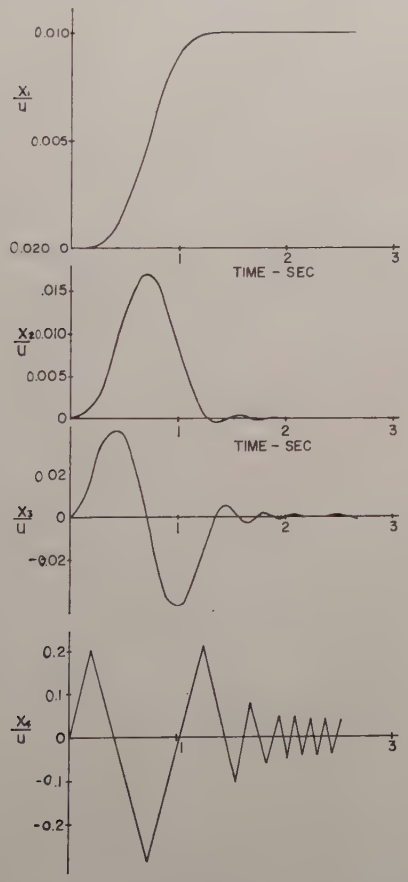


Fig. 5—Response to a step input of $x_1/u = 0.01$.

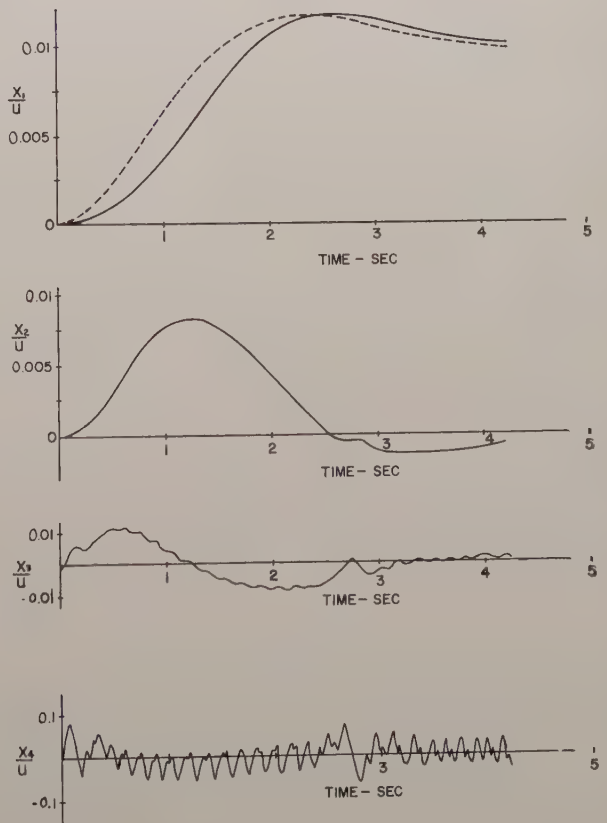


Fig. 6—Response of time-optimal system to second-order model input—no prediction of input.

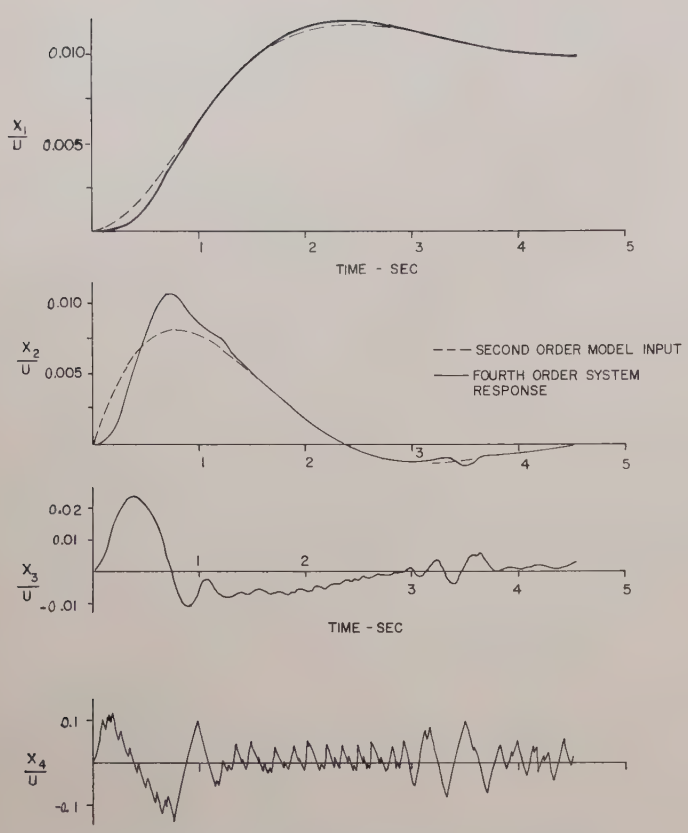


Fig. 7—Response of time-optimal system to second-order model input—first-derivative prediction of input at T.

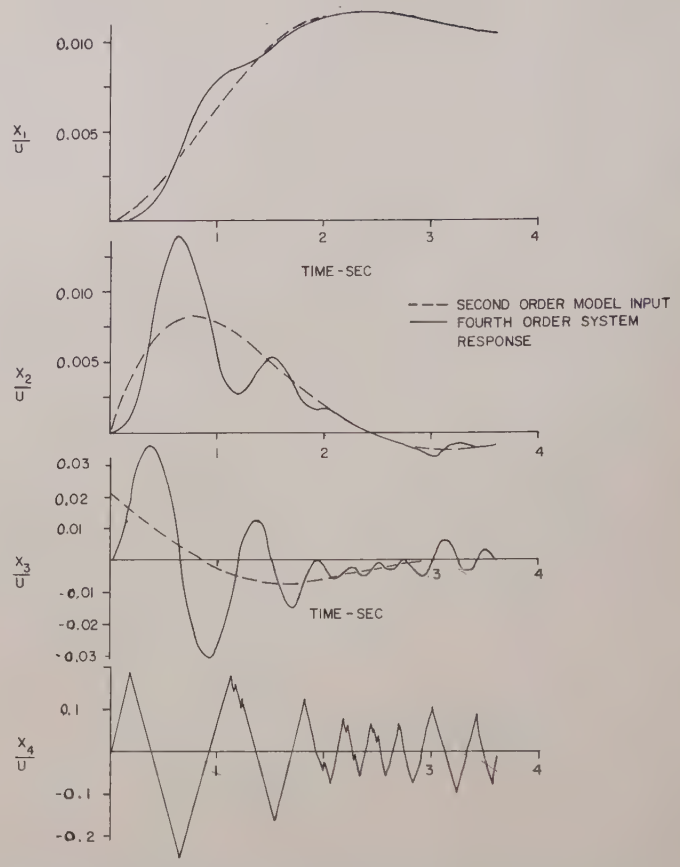


Fig. 8—Response of time-optimal system to second-order model input—first- and second-derivative prediction of input at T.

tions. Since the system is then being driven to the present value of the input, and this is changing quite rapidly during the first three seconds, there is a considerable lag in response. A high-frequency chatter of the relay is evidenced in the x_4 trace. Only the gross over-all motion is shown in x_4. There is a much higher frequency, a function of the switching error, which does not show on this scale.

Fig. 7 is the response to the same input, but with a first-derivative prediction term included. Following is much better and the very high frequency chatter is reduced.

Fig. 8 shows the response when both the first and second derivatives are used for prediction. It is felt that the step initial condition in x_3/u may be the cause of the overshoot. This is not too surprising, since the system is trying to catch up in minimum time and is not affected by what happens in between.

Conclusion

It is felt that the work presented here shows that time-optimal control of high-order systems with real roots is definitely feasible. For a digital system, limitations on the order of systems possible will be due to computer speed. The smaller the relative size of outside influences, the slower can be the computer. It is a relatively easy matter to include the derivatives of the input in the optimum switching solution. Thus the control problem, in addition to the regulator problem, may be solved. It has been shown that inclusion of these prediction terms gives markedly better following.

References

[1] "A Study to Determine an Automatic Flight Control Configuration to Provide a Stability Augmentation Capability for a High-Performance Supersonic Aircraft," Minneapolis-Honeywell Regulator Co., Minneapolis, Minn., Aero Rept. 48312-Final, Wright Air Dev. Center, Wright Patterson AFB, Dayton, Ohio, WADC Tech. Rept. 57–349; May, 1958.

[2] A. M. Hopkin, "A phase plane approach to the compensation of saturating servomechanisms," *Trans. AIEE*, vol. 70, pt. I, pp. 631–639; 1951.

[3] D. C. McDonald, "Nonlinear techniques for improving servo performance," *Proc. NEC* vol. 6, pp. 400–421; 1950.

[4] D. W. Bushaw, "Differential Equations With a Discontinuous Forcing Term," Stevens Inst. Tech., Hoboken, N. J., Experimental Towing Tank Rept. No. 469; January, 1953.

[5] J. P. LaSalle, "Basic principle of the bang-bang servo," *Bull. Am. Math. Soc.*, vol. 60, pp. 154; 1954.

[6] R. Bellman, I. Glicksberg and O. Gross, "On the bang-bang control problem," *Quart. Appl. Math.*, vol. 14, pp. 11–18; April 1956.

[7] R. W. Bass, "Equivalent linearization of nonlinear circuit synthesis and the stabilization and optimization of control systems," *Proc. Symp. on Nonlinear Circuit Analysis*, Polytech. Inst. Brooklyn, Brooklyn, N. Y., vol. 6, pp. 163–198; 1956.

[8] L. S. Pontriagin, V. G. Boltianskii, and R. Gamkrelidze, "On the theory of optimal processes," *Dokl. Akad. Nauk. SSSR*, vol. 110, pp. 7–10; September, 1956.

[9] R. Gamkrelidze, "Theory of time optimal processes for linear system." *Izvestia Akad. Nauk. SSSR*, Seria Matem, no. 4(22), pp. 449–474; 1958.

[10] V. G. Boltianskii, "Principle of the maximum in the theory of optimal process," *Dokl. Akad. Nauk, SSSR*, vol. 119, no. 6, pp. 1070–1073; 1958.

[11] L. T. Rozonoer, "L. S. Pontriagin's maximum principle in the theory of optimum systems," *Avtomat. i Telemekh.*, vol. 20, pp. 1320–1334; October, 1959.

[12] H. G. Doll and T. M. Stout, "Design and analog-computer analysis of an optimum third-order nonlinear servomechanism," *Trans. ASME*, pp. 513–525; April, 1957.
[13] I. Bogner and L. F. Kazda, "An investigation of the switching criteria for higher order servomechanisms," *Trans. AIEE*, vol. 73, pt. II, pp. 118–127; 1954.
[14] A. M. Hopkin and M. Iwama, "A study of a prediction-type air frame controller designed by phase space analysis," *Trans. AIEE*, vol. 23, pt. II, pp. 1–9; 1956.
[15] A. M. Hopkin and M. Iwama, "A Study of a Digitally Programmed Optimum Relay Servomechanism for Nonlinear Control of an Aircraft," University of California, Berkeley, Inst. Engrg. Res., Ser. No. 60, Issue No. 177, Rept. No. AD144448; February, 1957.
[16] F. Kurzweil, "The Analysis and Synthesis of Nonlinear Continuous and Sampled Data Systems Involving Saturation," Stanford Electronics Lab., Stanford, Calif., Tech. Rept. No. 2101–1; November 1959.
[17] E. N. Rozenwasser, "On the reduction of equations of nonlinear regulation systems to the simplest form," *Avtom. i Telemekh.*, vol. 21, pp. 15–19; January, 1960.
[18] C. R. Stone, C. W. Johnson, F. B. Smith, E. B. Lee, and C. A. Harvey, "Time Optimal Control of Linear Systens," Minneapolis-Honeywell Regulator Co., Minneapolis, Minn., Tech. Rept. R-ED 6134; September, 1959.
[19] E. B. Lee, "Mathematical aspects of the synthesis of linear, minimum response-time controllers," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 283–289; September, 1960.
[20] A. M. Hopkin and P. K. C. Wang, "A relay-type feedback control system designed for random inputs," AIEE, Paper No. 59–219; December, 1958.

Discussion

F. B. Tuteur (Yale University, New Haven, Conn.)

This paper presents an analytical procedure for obtaining the switching times in a high-order relay servo. The analytical solution for the switching times must in practice be implemented by a digital computer, but since the previous approach of using graphical (*i.e.*, phase plane or phase space) switching boundaries could be used easily only with second- or third-order systems, the present approach is a definite contribution to the theory of optimum switched systems.

There is, of course, some question in all systems of this type as to whether the complexity of the optimum controller is really worth the extra performance that is obtainable. Thus, in a second-order system the improvement in rise time, etc., which is obtainable from a switched system is usually no more than about 50 per cent better than that of the linear system, particularly when inputs are sufficiently large to saturate parts of the "linear," system. In the present instance, the author required an IBM-650 computer to control a relatively simple system. Although a special-purpose computer could probably be used, the computations are quite complex, and it seems that even a special-purpose computer would still be quite large and bulky, and therefore not very practical. I would be interested in comments by the author on this point.

Author's Comments: Dr. Tuteur raises the very valid and significant question of the practicality of optimal controllers—whether their advantages and improvements over conventional linear systems justify the additional complexity. It is, of course, a question which cannot be definitely answered in general at the present time. Theoretical work on optimal systems is relatively new and mechanization of such systems is still in its infancy. (Time-optimal control, on which this paper concentrates, is a special case of the general optimal control problem.) It is certain that a conclusion on their usefulness cannot be reached until tentative mechanizations are obtained and advantages of the systems are established. This paper is a step in that direction.

In spite of their infancy, however, I think it is possible to conjecture a bit on the potential usefulness of optimal controllers. It seems likely that they will always be more complex than the present-day linear controllers; in cases where the latter do an adequate job, the former will be of use mainly as a goal towards which to shoot, or as a yardstick for comparison. It is in applications where conventional techniques leave something to be desired that optimum or approximations to optimum control will be of practical value. Such problems as rational design of controllers for high-order systems, multivariable control, control with constraints to be enforced on some of the variables, control of systems with several independent but strongly coupled forcing functions, control of nonlinear plants, and adaptive control are all areas where additional complexity may be necessary to get the job done properly, if at all. Optimal control theory offers a rational and promising approach to these problems.

Control System Performance Measures: Past, Present, and Future*

W. C. SCHULTZ†, MEMBER, IRE, AND V. C. RIDEOUT‡, FELLOW, IRE

Summary—An increased amount of emphasis on the mathematical formulation of control system performance can be found in recent literature on automatic control. There are two areas of control system theory in which the application of performance measures is of interest: 1) the evaluation of control system designs in general, and 2) the design of adaptive control systems. In the former case, the performance measure is becoming an increasingly important aid to the control system designer. In the latter case, the performance measure takes on even greater significance, since adaptive systems include, by definition, a performance measure as an essential function which permits correction of system dynamic response during actual operation. Furthermore, the over-all evaluation of the adaptive loop itself presents new problems in the choice and use of performance criteria.

In the past, emphasis has been placed on various types of integrals, such as integral of error-squared and integral of the product of time and absolute error (ITAE); present emphasis is being placed on forms of integrals of a more general type; the trend for future emphasis appears to be in applications of statistical concepts and in attacking the problem of choice of the error measure in the adaptive system.

I. Performance Measures: Past

CONTROL system theory has been greatly enlarged during the course of the last two decades. Many tools for the analysis and synthesis of such systems have been developed and expanded. These include means for determining whether or not a system is stable, an important characteristic of any control system. Many criteria have been presented, such as those of Routh, Hurwitz, and Nyquist, which enable the linear servo designer to give a yes or no answer to the question, "Is the system stable?"

There are various kinds of stability, at least in nonlinear systems. However, stability alone, although a necessary requirement of a good system design, does not guarantee a suitable or usable system design. For example, specifications of accuracy and speed of performance must be met. Questions concerning performmance are more difficult to answer with a simple yes or no.

The problem of obtaining the answer to such a question is one of obtaining one number, or a limited group of such numbers, to describe the whole error, and then accepting or rejecting a system design on the basis of this number. In a mathematical sense, this number is a metric, since it gives the "distance" between two functions. The actual process of selecting which metric is to be used to "measure the distance" between a desired output time-function and an approximation to the desired function is, of course, the major difficulty. It is here that a certain amount of personal opinion is found. Surely the choice should be governed by usefulness, in that the particular metric which is selected must be a convenient one to use, as well as one which yields practical results.

Because real command signals, disturbances, and even system parameter values have random characteristics, it might be expected that choice of metrics or performance measures might be based upon statistical descriptions of these input functions. In the past, however, such a basis for metric choice seems often to have been more intuitive than explicit, and considerations of mathematical or computer equipment have often hidden a recognition of the underlying statistical problem.

The written history of performance measures dates back to 1942, the earliest date of any published material found by the authors. The classified nature of the development of servomechanism theory during World War II contributed to the delay in publication. For example, an early paper to appear in this country on the subject is the one written by Hall [1] in 1943, which had the classification, "restricted."

Although it was not stated in an explicit mathematical form of a metric, the first proposal of a measure of the error of a control system is the "deviation area" concept of Obradovic [2]. His paper appeared in a German publication in 1942. According to Obradovic, the starting point for obtaining the deviation area is to write the differential equation that describes the system, for example,

$$a_0 \frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_n x = 0. \tag{1}$$

An integration results in

$$0 = a_0 \frac{d^{n-1} x}{dt^{n-1}}\bigg|_0^\infty + a_1 \frac{d^{n-2} x}{dt^{n-2}}\bigg|_0^\infty + \cdots + a_{n-1} x\bigg|_0^\infty + a_n F. \tag{2}$$

* Received by the PGAC, May 26, 1960; revised manuscript received, October 31, 1960. Portions of this paper are reported in a Ph.D. dissertation supervised by Prof. Rideout. Other portions were sponsored by Cornell Aeronautical Laboratory, The University of Wisconsin Alumni Research Foundation, and the National Science Foundation.

† Electronics Dept., Cornell Aeronautical Lab., Inc., Buffalo, N. Y.
‡ Dept. of Elec. Engrg., and Math. Res. Center of the U. S. Army, University of Wisconsin, Madison, Wis.

It is now possible to solve for F, as shown below, and to plot it vs time, as shown in Fig. 1.

Fig. 1—Illustration of measure of error of Obradovic.

$$F = \frac{1}{a_n}\left\{a_0 \left.\frac{d^{n-1}x}{dt^{n-1}}\right|_{x=0} + a_1 \left.\frac{d^{n-2}x}{dt^{n-2}}\right|_{x=0} + \cdots + a_{n-1}x\Big|_{x=0}\right\}. \quad (3)$$

A good system, then, is typified by a small area F. Although it is not expressly stated as such, F may be called a metric, since it is a *number* which represents the "error" of the system.

The mathematical basis of the theory of measure of error, as it has developed, was set forth by Wiener [3]. He speaks of "operators" on functions, and the choice of the "best" operator. It is pointed out that one possible definition of "best" involves the assignment of a certain quantity as the error of performance of an operator, after which a particular operator is chosen from admissible operators in such a way that the error produced by the particular operator is as small as possible. The concept is applied in the ensuing example, where the difference between two functions F and g is to be minimized. It is then required that the following integral be a minimum.

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} [F(t) - g(t)]^2 dt. \quad (4)$$

Therefore, an integral is established as a measure of the error, and the system which causes this integral to be a minimum is called the "best" system.

The first application of the specific use of an integral as a measure of the error of a servo is found in the paper by Hall [1]. The actual error of a system is defined by Hall as the difference between the input and output of the system,

$$e(t) = r(t) - c(t), \quad (5)$$

where $e(t)$ is the error, $r(t)$ is the input and $c(t)$ is the output. Fig. 2 shows this relationship graphically, for a step input. Eq. (5) describes a function which serves as a statement of system error. However, as pointed out above, a function cannot readily serve as a metric, since it is not a number, but a collection of numbers. Furthermore, a function does not have certain properties which characterize a metric, namely:

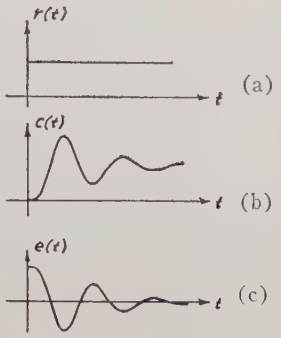


Fig. 2—Step input, output, and error.

1) that it is always positive or zero,
2) that it is zero if and only if the distance it represents is zero,
3) that it is a numeric.

The metric proposed by Hall is defined by

$$E = \int_0^\infty e^2(t)dt, \quad (6)$$

the integral of square of error (ISE). An inspection of (6) shows that it does have the three properties of a metric, in that it is a numeric, it can take on only positive values or the value zero, and it is zero if and only if $e(t)$ is identically zero.

One limitation of this metric, as applied to servos, is immediately apparent. The error, as given by (5), must be such that

$$\lim_{t\to\infty} e(t) = 0, \quad (7)$$

or the integral loses its meaning, unless the integral is truncated. This of course means that only systems which have zero steady-state error can be considered. Another difficulty, which is borne out by experience in applying the metric, is that it will not always lead to a practical system (see [64]). On the other hand, it is easily adapted to direct measurement by means of ordinary instruments, since many standard measuring devices give a direct reading that is proportional to the square of the input. Another feature is the mathematical convenience of squared error, since the integral of the squared error (ISE) often leads to forms that are well suited for computional purposes. This simple fact has often led to the use of criteria based on the square of error even though the use of other metrics would have led to more useful criteria. It was only after many such applications of the use of error squared measure that mathematical justification for the use of this convenient measure appeared [4]–[6]. These authors have shown that for a Gaussian stationary random input, the same "optimum" filter is obtained for any error measure chosen from a class of symmetrical functions.

Another quantity which has been proposed as a measure of error is the integral of error,

$$E = \int_0^\infty e(t)dt. \tag{8}$$

This is a more explicit form of the "deviation area" concept discussed by Obradovic, and mentioned above. It has also been discussed by Oldenbourg and Sartorious [7], Mack [8], Stout [9], and briefly by Laning and Battin [10]. This particular measure does not satisfy completely the properties of a metric, since this integral does not necessarily have to be positive. It may also be zero, even though the error itself is not zero. Nevertheless, in some cases, it may be considered to be a metric, and it has been used to describe system performance. It does also have the advantage, as the integral of square-of-error has, that standard instruments exist to enable direct measurements of the value of the integral. It is pointed out by Stout [9] that in certain cases the integral is very simply evaluated, in the following manner. If $r(t)$ is a unit step, and if the Laplace transform of the ratio of output to input can be expressed by

$$\frac{C(s)}{R(s)} = \frac{(\tau_1 s + 1)(\tau_2 s + 1) \cdots (\tau_m s + 1)}{(T_1 s + 1)(T_2 s + 1) \cdots (T_m s + 1)}, \tag{9}$$

then the value of the integral, called the "control area," is

$$\begin{aligned} E &= \int_0^\infty e(t)dt \\ &= (T_1 + T_2 + T_3 + \cdots + T_m) \\ &\quad - (\tau_1 + \tau_2 + \cdots + \tau_m). \end{aligned} \tag{10}$$

The proof is readily obtained by the use of the Laplace transform and the associated final-value theorem. Therefore, in cases where applicable, the integral of error is a very convenient measure to use.

In a paper by Nims [11] a discussion of "control area" is presented, which is illustrated in Fig. 3. Nims discusses a disadvantage of the use of control area as a measure of error. This disadvantage is that a system which has an oscillatory response has both positive and negative control area. The negative area therefore subtracts from the final value of the integral, which causes the integral to yield a value that is not truly descriptive of the over-all error. Nims therefore proposed a "weighted control area," as described by

$$E = \int_0^\infty te(t)dt. \tag{11}$$

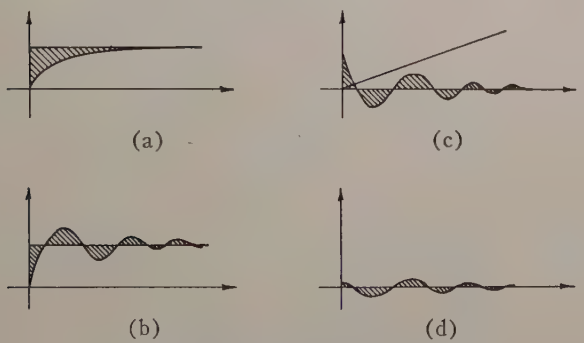


Fig. 3—Control area and weighted control area. (a)–(c) Control area. (d) Weighted control area.

This integral is illustrated in Fig. 3(c) and 3(d). An advantage of the use of the weighting function is that the large initial error (for step inputs) does not give rise to such a large contribution to the value of the integral. This consequence of the use of a weighting function is a desirable one, since the initial error in an ordinary servo is due to the inertia of the system, and hence, largely unavoidable. Therefore, a metric which is relatively insensitive to the initial error is of interest.

The integral of the absolute value of the error has been proposed as a measure of system error. This integral,

$$E = \int_0^\infty | e(t) | dt, \tag{12}$$

has been used in connection with analog computer solutions of servo problems by Fickeisen and Stout [12] and also by Caldwell and Rideout [13]. It is interesting to note that this metric first appeared in connection with analog studies. This is not surprising however, since it is one of the simplest of all metrics to use with analog computers, though usually difficult to deal with analytically. This metric is also discussed by Laning and Battin [10].

A paper by Graham and Lathrop [14] created a great deal of interest. Their paper presents as a metric the integral of time multiplied by absolute error (ITAE). It is defined as

$$E = \int_0^\infty t | e(t) | dt. \tag{13}$$

As in (11), the time weighting serves to reduce the contribution of the large initial error to the value of the integral, as well as to place an emphasis on "late" errors. Another important feature of the Graham and Lathrop paper is the detailed discussion and presentation of results of a comparison of the various metrics with the ITAE metric. For example, it is shown that the ITAE alone has the selectivity needed for a good metric; that is, a minimum value of the integral is readily discernible, as system parameters are varied. The other metrics, integral of error, integral of absolute error, and integral of error squared, do not have this same property of selectivity, especially for higher-order systems. In addition to the comparison with the known metrics, mentioned above, results of a study of other measures, *i.e.*,

$$E = \int_0^\infty te^2(t)dt, \tag{14}$$

$$E = \int_0^\infty t^2e^2(t)dt, \tag{15}$$

$$E = \int_0^\infty t^2 \mid e(t) \mid dt, \tag{16}$$

are given. These metrics have a selectivity which is desirable, but are given no further consideration because of the increased difficulty in handling them, either analytically or on an analog computer. The conclusion drawn from the Graham-Lathrop paper is therefore as follows: the integral of ITAE is a superior metric to use as a basis for a criterion because of its selectivity and relative ease of calculation. As a summary of results, tables of standard forms are given for the ITAE, as compared with the binomial standard forms and the Butterworth standard forms. Companions of these tables are the corresponding transient response curves, for step input. In closing the discussion of this paper, it might be remarked that certain nonlinear systems are also "optimized" by use of the ITAE.

sults. This philosophy is reflected in the following remarks:

> . . . The ultimate decision of what constitutes good performance is based on human judgment or even personal opinion . . . the end result is in the nature of a hit or a miss. . . . It is not possible in a general way, to legislate for all cases, and the field must be narrowed.

It is interesting to note that the concept of moments-of-error-squared can also be used to describe two other metrics discussed earlier. Eq. (6), the integral of error squared, can be said to be the zeroth moment of error-squared. Eq. (14) is the first moment of error-squared, while (15) is the second moment of error-squared.

Wescott also speaks of "special penalty cases," where "catastrophe" results if, for example, more than 10 per cent overshoot occurs. Such cases are not considered in his paper.

One approach to the problem of considering special penalty cases is suggested by Whiteley [16]. A table of coefficient values of "standard forms" for optimum systems based simply on a maximum amount of overshoot is presented. These coefficients are listed in Table I. The table shows, for example, that the parameter for a second-order, "class A" system, should be equal to 1.4.

TABLE I

WHITELEY'S STANDARD FORMS

Class of System	$Q=\frac{\theta_0}{\theta_i}$	Basis	Maximum Per cent Overshoot	Coefficients
A Zero Displacement Error	$\frac{\omega_0^n}{p^n+a_1p^{n-1}+\cdots+\omega_0^n}$	$Q=1$ for frequencies up to $f\cong f_1$, where $f_1=\frac{0.8\omega_0}{2\pi}$	5 per cent 8 per cent 10 per cent	1 2.6 3.4 2.6 1 1 2 2 1 1 1.4 1
B Zero Velocity Error	$\frac{2p+\omega_0^n}{p^n+\cdots 2p+\omega_0^n}$	Maximum overshoot of 10 per cent, and no subsequent undershoot	10 per cent 10 per cent 10 per cent 10 per cent 10 per cent	1 11 43 83 73 25 1 1 9 29 38 18 1 1 7.2 16 12 1 1 5.1 6.3 1 1 2.5 1
C Zero Acceleration Error	$\frac{yp^2+2p+\omega_0^n}{p^n+\cdots+yp^2+2p+\omega_0^n}$	Maximum overshoot as given, with one small undershoot	10 per cent 15 per cent 20 per cent 20 per cent	1 36 251 485 251 36 1 1 18 69 69 18 1 1 7.9 15 7.9 1 1 6.7 6.7 1

About the same time that the Graham and Lathrop paper appeared, a paper concerning another integral form was presented by Wescott [15] in England. This integral, called the "moment-of-error-squared" is

$$E = \int_0^\infty te^2(t)dt. \tag{17}$$

The criterion based on this metric is called the "minimum-moment-of-error-squared criterion." Comparisons are made with the integral-of-error-squared criterion, and cases where the weighted integral has the advantage over the unweighted integral are pointed out. On the other hand, for some cases Wescott's measure leads to the same difficulty as the ISE in giving impractical results. The transient response for a step input will then be such that the maximum overshoot is less than five per cent. Thus, the special case of specifying a maximum amount of overshoot is an early attempt to legislate against the "catastrophe" cases. Standard forms for the ITAE measure have also been given by Graham and Lathrop [14].

In their book on servomechanisms, James, Nichols and Phillips [17] discuss in some detail the minimization procedures of Wiener, as based on the root-mean-square measure. Some attention is also given in this book to the work of Hall and the ISE. It is pointed out that Hall's method is a convenient mathematical tool and should therefore be used whenever applicable. It is also

pointed out that certain disadvantages exist (as mentioned above), and that consequently other error-weighting devices might be employed. An illustration is given of an error vs weight-of-error curve (see Fig. 4.)

In a discussion of control system design techniques, Truxal [18] describes the minimization process in detail. In presenting some disadvantages of the integral of square of error as a metric, the use of "error vs importance-of-error" curves is suggested. For example, if the system error is small, motor torque should be small; or if system error is large, full torque should be applied. This type of error vs importance-of-error curve is illustrated in Fig. 5.

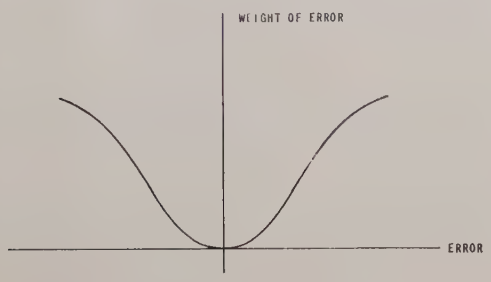


Fig. 4—Weight-of-error curve of James, Nichols, and Phillips.

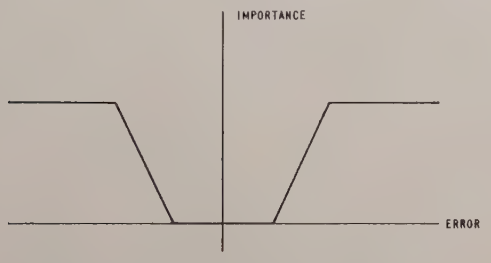


Fig. 5—Truxal's importance-of-error curve.

Gibson [19] has presented a paper which summarizes a number of methods of evaluating system performance. In this paper, a simple control system is analyzed by the several methods, including the characteristic plots of Nyquist, Bode, and the Evans root-locus. In addition, an integral metric is employed and examined, the ITAE of Graham and Lathrop. A table summarizes the results of applying each of the methods, by showing what information about system performance is obtained from each method.

A generalization and extension of the type of measures discussed thus far was presented in a previous paper [20]. This form provides a versatile measure for servo performance, one that can be adapted to fit the many situations that arise in servo problems. Eq. (18) states the generalization

$$E = \int_0^\infty F[e(t), t]dt. \tag{18}$$

The versatility of this form of error measure is inherent in the interpretation and formation of the F functions, which can be either functions of magnitude of error alone, $F[e(t)]$, or functions of magnitude and time-weighting functions. For example, F functions could be chosen as any of the forms discussed earlier, such as $e(t)$, $e^2(t)$, $|e(t)|$, $t|e(t)|$, and $te^2(t)$.

If a geometrical interpretation is made for the F functions, a great deal of insight can be gained for adapting (18) to special cases, particularly for "catastrophe" cases. An illustration for $F[e(t), t] = |e(t)|$ is shown in Fig. 6. Upon observing the F-function characteristics, it is noted that for positive inputs, the left quadrant represents overshoot-error weighting, while the right quadrant represents undershoot-error weighting. This interpretation opens the way to anti-overshoot measures, as illustrated in Figs. 7 and 8. If a certain small amplitude of error is permissible (even in the steady state) an F function of the type shown in Fig. 9 could be used. It is apparent that almost an infinity of F functions could be constructed, to be applied to special solutions. The optimum step responses, relative to the F functions discussed in this paragraph, are shown in Fig. 10.

Thus far all discussion has centered on a definition of error given as input minus output [see (5)]. It is possible to extend this definition of error to include a more general interpretation of error. Thus a model error or model-system error may be defined. The basis for this extension is shown in Fig. 11. If the "model" is considered to have a unity transfer function, the error is the same as that described in (5). The basis for a more complex interpretation of system error is provided by choosing a "desired transfer function" for the model and comparing its output with the actual system output. This scheme is discussed in a number of papers, in connection with adaptive control system design.

Performance measures of a form somewhat different from those discussed above can be derived if the "model" of Fig. 11 is a pure delay. System error is then the delay-error, discussed by the authors in a previous paper [21]. Delay-error for a step input is illustrated in Fig. 12.

Integrals of this different form have been discussed by a number of authors. For example, a measure of amplifier distortion was described by Aigrain and Williams [22]. This integral, called the "transient distortion," is defined as

$$I = \int_0^\infty [f_2(t) - f_1(t_1 - T_0)]^2 dt, \tag{19}$$

where $f_1(t)$ is a step function input, $f_2(t)$ is the amplifier transient output and $f_1(t - T_0)$ is the input delayed by an amount T_0. The paper is devoted mainly to a method of evaluating the integral, and the application is limited to the study of amplifiers.

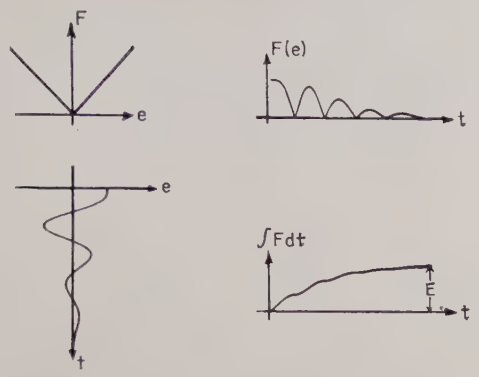


Fig. 6—Integral of absolute error metric.

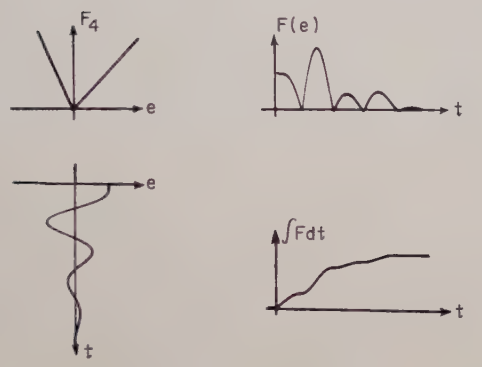


Fig. 7—Anti-overshoot metric.

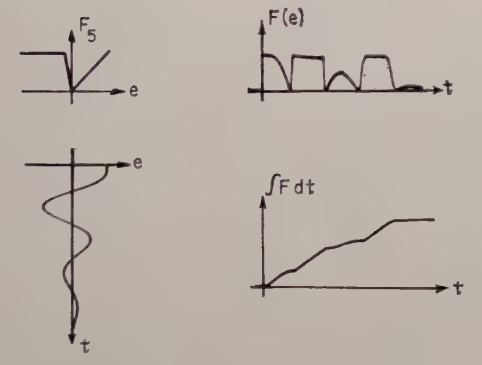


Fig. 8—Anti-overshoot metric.

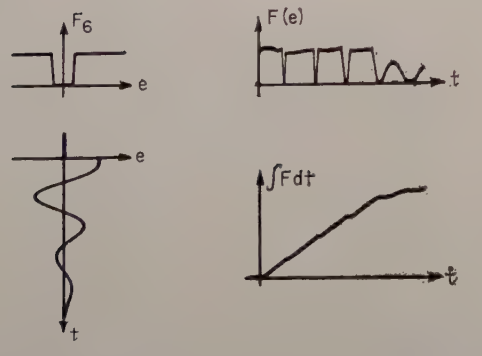


Fig. 9—Special measure which permits small errors.

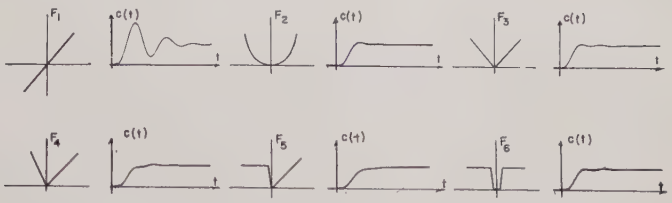


Fig. 10—Optimum response to step for various measures.

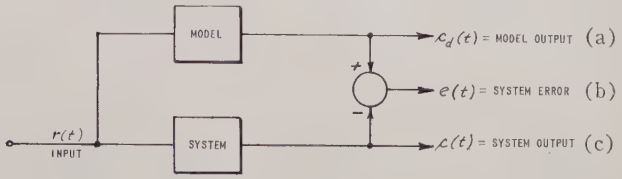


Fig. 11—(a) Model transfer function $\equiv 1$. $e(t) = r(t) - c(t)$ = ordinary error.
(b) Model transfer function = pure delay τ. $e(t) = e(t,\tau) = r(t-\tau) - c(t)$ = delay error.
(c) Model transfer function = desired system. $e(t) = c_d(t) - c(t)$ = model comparison error.

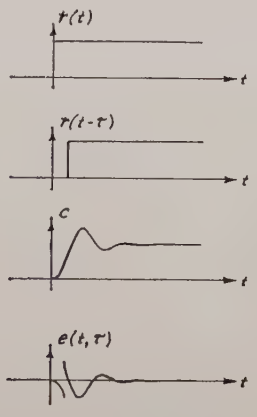


Fig. 12—Delay error.

Another proposal of an integral of this form has been made by Lee and Wiesner [23]. In their discussion of the application of correlation functions in determining an optimum linear system, the following error expression is introduced:

$$\epsilon = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} [f_0(t) - f(t-\alpha)]^2 dt, \qquad (20)$$

where $f_0(t)$ is the system output and $f_m(t)$ is the desired filter output. A delay α is permitted in the expression of the output. Consequently, correlation functions are introduced, as an expansion and examination of (20) shows.

More recently, Spooner and Rideout [24] discussed this form of error measure, for systems having stationary random inputs. They have called this function the generalized error function (GEF). This function is defined as

$$\text{GEF} \equiv E(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} [r(t-\tau) - c(t)]^2 dt. \qquad (21)$$

In a study that closely followed this work, the authors [21] considered a similar form, but for transient inputs. This form is given by

$$E(\tau) = \int_{-\infty}^{\infty} [r(t-\tau) - c(t)]^2 dt. \tag{22}$$

An interesting comparison of (21) and (22) can be made [24]. An expansion of (21) results in

$$E_r(\tau) = \phi_{11}(0) + \phi_{22}(0) - 2\phi_{12}(\tau) \tag{23}$$

for the random input signal case. The ϕ's are input, output, and cross-correlation functions. For the transient input signal case,

$$E_t(\tau) = \psi_{11}(0) + \psi_{22}(0) - 2\psi_{12}(\tau). \tag{24}$$

The ψ's are the input, output, and cross-translation functions. (Translation functions were introduced by Newton [25].) The process of measuring translation functions on an analog computer is simpler than that of measuring correlation functions, since this type of transient signal is the response to a step input, which can easily be delayed without a delay line. Thus, the analysis by use of this transient analog provides a means of simplified mechanization, when it can be applied.

An illustration is given in Fig. 13. The upper system represents a linear network L_2 with a random signal input $r_r(t)$. This input can be considered to be the output of a linear filter L_1, which has a white noise input. Associated with this system are the input- and output-auto-correlation functions, and the respective *power* spectral densities. $E_r(\tau)$ for this system is given by (23). Fig. 7(b) shows the transient analog of the random-input case. For this system, the input is $r_t(t)$, the impulse response of a particular linear filter. Associated with this system are the input- and output-auto-translation functions, and the respective *energy* spectral densities. $E_t(\tau)$ for this system is given by (24). For the case of a linear L_2 network, $E_r(\tau)$ and $E_t(\tau)$ are numerically equivalent, provided that $r_t(t)$ satisfies

$$r_t(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_{11}^{+}(j\omega)\epsilon^{j\omega t} d\omega, \tag{25}$$

where $\Phi_{11}(j\omega) = \Phi_{11}^{+}(j\omega)\Phi_{11}^{-}(j\omega)$, and $\Phi_{11}^{+}(j\omega)$ has only upper-half plane poles. For an illustrative example, consider

$$\Phi_{11}(j\omega) = \frac{1}{1+\omega^2}; \tag{26}$$

then

$$\Phi_{11}^{+}(j\omega) = \frac{1}{1+j\omega}$$

and

$$r_t(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\epsilon^{j\omega t}}{1+j\omega} d\omega$$
$$= \epsilon^{-t}, \quad t > 0$$
$$= 0, \quad t < 0. \tag{27}$$

Thus, the linear system L_2 having the input signal with the power spectrum of (26) can be represented on an analog computer by the transient analog having an input $r_t(t)$ given by (27). The inter-relationships among these various quantities are summarized [26] in Fig. 14.

Performance measures "past" may be appropriately summarized by the forms shown in (18). The early performance measures, such as integral of error squared and integral of absolute error, are included in these forms, as are the time-weighted forms. In addition, there is provided a means for treating certain specialized performance characteristics. To a great extent, all these

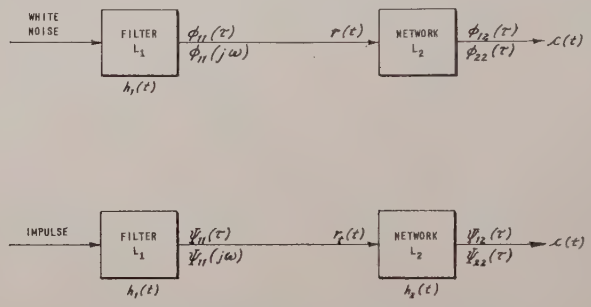


Fig. 13—Transient analog for random signal case.

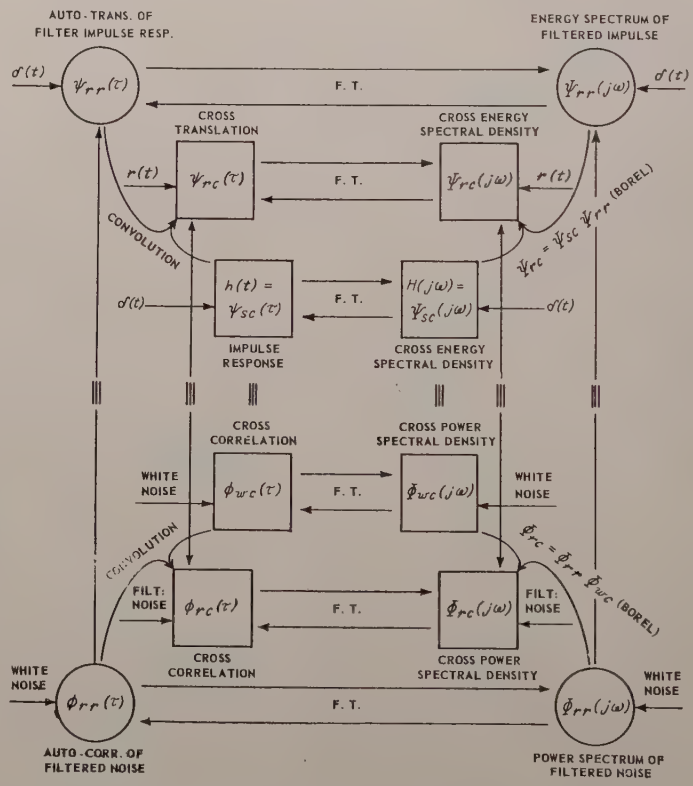


Fig. 14—Interrelationships of system characteristics.

forms have been used only for step inputs. It is also apparent that these forms are based more on intuitive than on mathematical concepts. Nevertheless, the development of these forms represents a significant step in the evolution of performance measures.

II. Performance Measures: Present

For the purposes of discussion in this section, the "present" will be somewhat loosely interpreted as including the past year or two. During this period, the subject of performance measure in control system design has received increased emphasis in two areas:

1) the evaluation of control systems, where performance measures receive the emphasis once given to stability study,
2) the use of performance measuring devices as essential integral parts of self-adaptive control systems.

Approaches to the performance measure problem include the following two major areas of mathematical theory:

1) statistical and probabilistic methods, originated by Wiener,
2) methods based on the calculus of variations and extensions of this theory by Bellman (dynamic programming).

Statistical and Probabilistic Concepts

Wiener's original work has been extended by others [27]–[33] but in general an estimate of the mean square error has been used as a performance measure. Kaufman [34] concerned himself with system optimization based on other than estimates of mean-square error, in linear systems with stationary (but not necessarily Gaussian) inputs. In particular, he has suggested measures of the form

$$E = c_2\overline{e^2} + c_4\overline{e^4} + c_6\overline{e^6} \tag{28}$$

for cases where the mean square alone is not sufficiently meaningful. For those cases where the input and disturbances are independent and each is symmetrical, he shows how to evaluate and minimize E analytically by use of approximations for the correlation functions which occur in the integrands.

A performance measure based on a probabilistic error has been proposed and described by Zaborszky and Diesel. This measure is defined as follows:

$$E = \int_0^\infty F[(e(t), t, v_1, \cdots, v_r]p(t)dt, \tag{29}$$

where

$e(t)$ = system error,
t = time,
v_i = system parameters,
$p(t)$ = probability distribution of the times at which the output is utilized.

Although the paper introduces some ambitious objectives and advances claims of broader usefulness and greater physical meaning for this measure than for existing criteria, substantiation of these claims is not apparent. In particular, reference is made to the need for "engineering judgment" in selecting a specific $F(e)$ function, and also to exclusive use of the commonly used $F(e) = e^2$.

Thus, (29) becomes in reality

$$E = \int_0^\infty e^2(t)p(t)dt. \tag{30}$$

Eq. (30) therefore represents the main theme of the proposed probabilistic measure, a form which is certainly contained in the class of measures defined by (18). It is also apparent that the $p(t)$ factor in (29) is in a sense redundant, since any form of measure that can be obtained from (29) can also be derived from (18).

Subsequent papers [65], [66] of these authors further emphasize the attention placed on the time-weighting function $p(t)$, without regard to the consideration of choice of the amplitude-weighting function $F(e)$. For cases where the use of (30) is purposeful, advantages can be gained by pre-programmed computer solutions for evaluation purposes, if the use of a large-scale digital computer is justified.

Some early applications of the use of correlation functions in performance measures were discussed earlier in this paper [23], [24]. Additional schemes are now summarized.

Reswick [37] discusses methods which make use of auto- and cross-correlation functions and provide means whereby the system impulse response can be obtained during system operation. A special device, called the delay-line-synthesizer (DLS), is used for impulse response determination. This device approximates the deconvolution process; *i.e.*, $h(t)$ is obtained from the deconvolution of the integral

$$\phi_{rc}(\tau) = \int_0^\infty h(u)\phi_{rr}(\tau - u)du, \tag{31}$$

since for white noise system input, $\phi_{rr}(\tau)$ is an impulse and $\phi_{rc}(\tau)$ becomes equal to the impulse response.

Another basic application of correlation techniques is introduced by Anderson, *et al.* [38], where a figure of merit called the impulse response area ratio (IRAR) is used. This performance measure is a key element of an adaptive control system discussed in the paper. Aseltine has proposed a definition of *adaptive control* in which a performance measure has an important role [39]. This definition consists of the following three elements:

1) continuous measurement of system dynamic performance, while the system is operating,
2) means of converting this measure of performance into a number that describes how good the performance is,
3) readjustment of system control parameters on the basis of 1) and 2).

Although differences of opinion may exist on the question of what does or what does not constitute an adaptive system, the above viewpoint that a performance measure vitally affects the operation and design of the (adaptive) control system is widely accepted.

The method of measuring the system performance adopted in the Anderson paper is one that is attributed to Lee [40] and is the one also used by Reswick. The method of evaluating the system performance departs from the usual case of minimizing or maximizing some kind of integral. Instead, a performance measure is selected that has the value zero for "optimum" conditions, and positive or negative values for departure from the optimum conditions. The measure is based on the damping ratio or relative stability of the system. The figure of merit F is differentiated with respect to the variable to be adjusted in the system transfer function, to obtain the expression for the "gain" of the adaptive loop. An extension of this concept is presented subsequently in this paper.

Moments of the system impulse response have been shown to provide a means for determining system dynamic characteristics [41]. More recently, Goodman and Hillsley [42] have described a method for obtaining such moments continuously.

The nth moment of the impulse response is defined by

$$H_n = \int_0^\infty t^n h(t)dt. \tag{32}$$

A similar definition can be given for moments of the correlation function. For example, the nth moment of the cross-correlation function is defined by

$$C_n = \int_{-\infty}^{\infty} \tau^n \phi_{rc}(\tau)d\tau. \tag{33}$$

Goodman and Hillsley have demonstrated a method for measuring the correlation function moments continuously on an analog computer. Thus, a means is provided for obtaining a measure related to impulse without actually measuring the impulse response. An extension of this concept is discussed subsequently in this paper.

Variational Calculus Techniques

An extension of the Wiener optimization technique using variational calculus was presented recently by Murphy and Bold [43]. In this discussion, an investigation of a square-of-error criterion with an arbitrary weighting function is presented. Their error criterion is defined by

$$E = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} W(t)e^2(t)dt, \tag{34}$$

where $e(t)$ is the system error and $W(t)$ is a function of time.

The minimization procedure is a method for obtaining the optimum system transfer function for an arbitrary $W(t)$. This use of $W(t)$ as a weighting function is more general than the performance measure of Zaborszky-Diesel, as indicated in (30), since the restrictions on $W(t)$ are not as great as those placed on the $p(t)$.

Dynamic programming is a relatively new functional equation technique introduced by Bellman [44]. The method extends the concepts of the calculus of variations, enabling analytical treatment of a wider class of problems. In general, solutions obtained by this method are obtained numerically, through the use of high-speed digital computers [45].

The investigations of Bellman and Kalaba [45], [46] have been directed toward the laying of the mathematical foundations of dynamic programming. Others [47]–[49] have investigated the application of these techniques to adaptive system design problems. Merriam [49] points out that the dynamic programming method permits an error minimization algorithm to be developed for an arbitrary error criterion. However, he goes on to a discussion of a computationally more practical subclass in which a quadratic error measurement is used.

In another approach to the problem of utilizing a general performance index, Kalman and Koepcke [50] have proposed a class of quadratic indices, as applied to continuous and sampled-data control systems. This form is defined by

$$J_N = \int_0^{NT} [z'(t)Qz(t) + \gamma m^{*2}(t)]\omega(t)dt, \tag{35}$$

valid for the time interval $0 \leq t \leq NT$; Q and $Z(t)$ are matrix representations of the state of the system, or a measure of the error at a time t; $m^*(t) = m(kt)$, a sequence of numbers that describes the control signals of the system; γ is a constant which weights the relative importance of minimization of errors, and the control energy associated with the $m^*(t)$ term; $\omega(t)$ is a time-weighting function, which is used to indicate the importance of errors occurring at various instants of time. This performance index is more general than the one suggested in (31), since $\omega(t)$ could be defined as a probability density function, or as some other function of time, for example, $\omega(t) = t$. Another interesting weighting function is the exponential function λ^t where λ is some constant. Such a choice leads to performance indexes such as

$$J = \int_0^T [e^2(t)]\lambda^t dt, \tag{36}$$

and

$$J = \int_0^T [e^2(t) + k\dot{e}^2(t)]\lambda^t dt. \tag{37}$$

It is pointed out by Kalman and Koepcke that certain difficulties evaluating system designs by use of these performance measures can be overcome by use of dynamic programming techniques.

A measure similar to (37) is described by Topitsyn [58] in a Russian paper. He discusses methods of optimization for a given maximum amount of overshoot.

III. Performance Measures: Future

The fact that performance measures are taking on increased importance in control-system design is evidenced by the vast amount of material that has appeared on the subject in recent years. A significant sampling of the investigations which have been made is summarized in the first two sections of this paper. A trend that is clearly implied, and, in some cases, pointed out directly, is one of an increased emphasis on the use of statistical and probabilistic concepts. Inherent in these approaches are more direct procedures for arriving at optimum designs than were once possible. Particularly with the aid of computers, designers are free to use a wide variety of performance measures. A good example is seen in some of the applications of dynamic programming methods.

A digression is in order at this point. In a sense, a motion has been made [51] and seconded [46] to re-evaluate what is meant by the term "optimum." This follows rather naturally, as control systems become more sophisticated and design procedures rely more and more on improved mathematical processes and computation techniques. A certain amount of arbitrariness remains, however, in the choice which can now be made from a wide variety of performance measures. This choice is much like the engineer's other design problems. Faced with the need to build a control system to perform some function, he must base his system design on some composite performance measure, which is influenced by considerations of cost, time and convenience, as well as by system performance. The need for an additional factor in a performance measure, that of reliability, is also becoming more apparent. The choice of the performance measure, by the engineer or his customer, is thus a second-order design problem and is based on some higher-order and more complex performance measure of the performance measure, as well as, again, on considerations of cost, time and convenience. Thus the ITAE measure, which is certainly convenient, is used because it has been shown to give systems whose response to an input step has little overshoot and fast time-of-rise. Thus, notions regarding overshoot are used to govern choice of a criterion which does not directly measure overshoot.

Another reason for choice of a criterion is to provide a simple means for comparison of quite different systems. Here again, simplicity is important in choice of a performance measure.

In going from a consideration of control system errors with step function inputs to errors with random inputs, the effect of the choice of performance measure, and its subsequent minimization, on statistical properties of the error becomes important. An example is shown in Fig. 15, for $f(\epsilon) = \epsilon^2$. The expected value of the function of error is

$$E[f(\epsilon)] = \int_{-\infty}^{\infty} f(\epsilon) p\{f(\epsilon)\} d(f(\epsilon))$$

$$= \int_{-\infty}^{\infty} f(\epsilon) p(\epsilon) d\epsilon. \tag{38}$$

Thus, the criterion output depends on the error probability density $p(\epsilon)$.

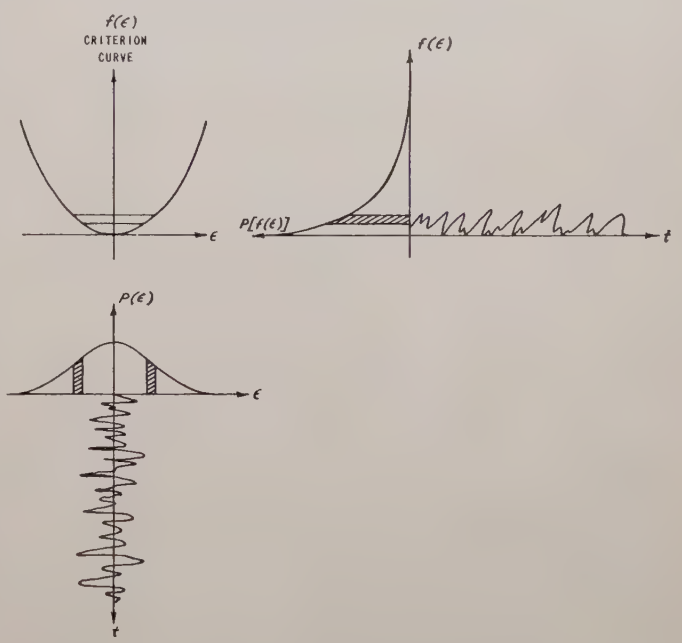


Fig. 15—Probability density curve and function of error curve.

If the system input $r(t)$ is Gaussian, the output $c(t)$, and thus the error $\epsilon(t)$, will also be Gaussian for a linear system. It has been shown [4]–[6] that within a wide class of functions the choice of $f(\epsilon)$ does not really matter, since the same optimum linear system results.

Thus, in a self-adaptive system based on a linear system with Gaussian input, the distribution of the output of the system cannot be altered by parameter adaptation. Consequently, only the mean and mean-square can be reduced.

However, if the input is not Gaussian, or if the system is nonlinear, the choice of the criterion curve will affect the system output probability distribution in addition to affecting the mean and mean-square. Thus, the idea of shaping the criterion-curve to provide the desired error-probability density curve, by adaptation or by hand, appears.

In applying this notion, it would seem desirable to measure some finite number of important moments, combine this information, and use it to control parameters in such a way as to minimize some weighted average of these moments. An example of such a process is shown in Fig. 16, where a symmetrical distribution of the error is assumed. Here estimates of $\overline{\epsilon^2}$, $\overline{\epsilon^4}$ and $\overline{\epsilon^6}$ are periodically assessed by a computer, and used to control parameters α, β and γ of a system, in an adaptive scheme. If a constant relative weighting of those moments is required, a single $F(\epsilon)=a\epsilon^2+b\epsilon^4+c\epsilon^6$ may be used.

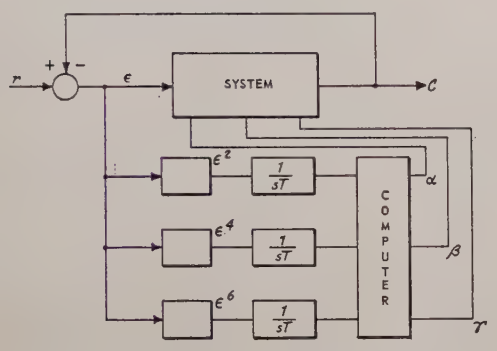


Fig. 16—Control of parameters by weighted average of moments of error.

These ideas are applicable to the case where the delay-error $r(t-\tau)-c(t)$ or prediction-error $r(t)-c(t-\tau)$ is of interest, rather than the ordinary error $r(t)-c(t)$, and mechanization of such error measurement merely requires the dead-time delay of the appropriate signal. In the existing body of theory that will be useful in developing these probabilistic ideas are the works of Baum [52] and Axelby [33] and others. Baum discusses limiter curves defined by the error integral, for Gaussian inputs.

$$y(x) = \frac{1}{K}\frac{1}{\sqrt{2\pi\sigma_{0s}}}\int_0^x \mathcal{E}^{-z^2/2\sigma_{0s}}dz. \tag{39}$$

The shape of the output probability density function can be varied by adjusting σ_{0s}, the rms value of the limiter curve. In particular, the output probability density curve can be made to be uniform, over the range $\pm\frac{1}{2}K$, if the rms value of the limiter curve equals the rms value of the input noise.

Axelby [33] demonstrates the effect of noise transmission through nonlinear elements, such as saturation, dead zone, and "bang-bang." The latter case is a limiting case of (40) for zero rms value of the limiter curve.

Still [53] describes an analog simulation of a probability filter. The concept introduced in this work is one of giving small weighting to large errors of low probability of occurrence. Thus, only the errors of high probability of occurrence will affect the system. An example using a nonlinear probability filter illustrates the effectiveness of this scheme.

Applications in Adaptive System Simulation: Digital

Extensions to the work of Goodman and Hillsley, and Aseltine, *et al.*, are discussed in a current paper on an adaptive system simulation [54]. This investigation considers a multi-dimensional figure-of-merit. An assumption is made that a desired or "ideal" system dynamic response can be characterized by a set of power moments of the system impulse response. A corresponding set of moments can be measured during system operation. Subtracting the desired moments provides a set of moment corrections, ΔH_1, ΔH_2, which are required to make the actual system response correspond to the desired response, insofar as is possible through correspondence of the first n moments. The actual moments are computed in terms of the system input-output cross-correlation, where the input is white noise.

Since the moments are functions of the adjustment parameters, the total differential of the nth moment is

$$dH_n \cong \Delta H_n = \frac{\partial H_n}{\partial x_1}\Delta x_1 + \frac{\partial H_n}{\partial x_2}\Delta x_2 + \frac{\partial H_n}{\partial x_3}\Delta x_3 + \cdots \tag{40}$$

where the x_i are the parameters of the system to be adjusted adaptively. The expression which forms the basis for simultaneous adjustment of all moments is the matrix expression

$$[\Delta H] = [A][\Delta x], \tag{41}$$

where $[\Delta H]$ is a moment-correction column matrix, $[\Delta x]$ is a parameter-change column matrix, and $[A]$ is a matrix whose elements are defined by

$$a_{nj} = \frac{\partial H_n}{\partial x_j}. \tag{42}$$

The parameter-change increments are

$$[\Delta x] = [A]^{-1}[\Delta H]. \tag{43}$$

An elaborate simulation on an IBM-704 was prepared for investigating this concept, and initial tests conducted with the simulation give evidence that the scheme is feasible. Further tests are currently planned, and are being made with the simulation.

A diagram of the over-all system is shown in Fig. 17.

Applications in Adaptive System Simulation: Analog

In another current paper [55], an analog computer study of an adaptive control system is described. Error criteria are discussed in terms of a parameter-perturbation adaptive system.

A block diagram of the system is shown in Fig. 18. In this system, if we can assume that the small parameter-perturbation signal α, sin ω, t causes a linear modulation of the error, then

$$\epsilon(t) = \epsilon_0(t)[1 + \alpha_1 m_1 \sin\omega_1 t], \tag{44}$$

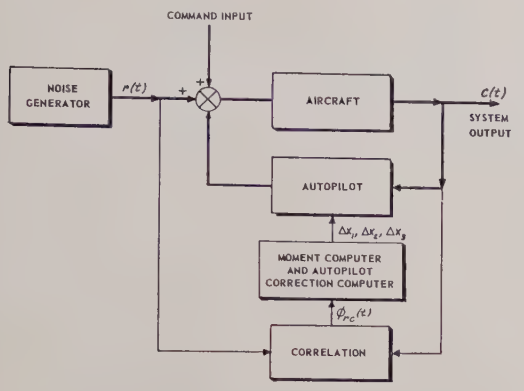


Fig. 17 —An adaptive flight-control system.

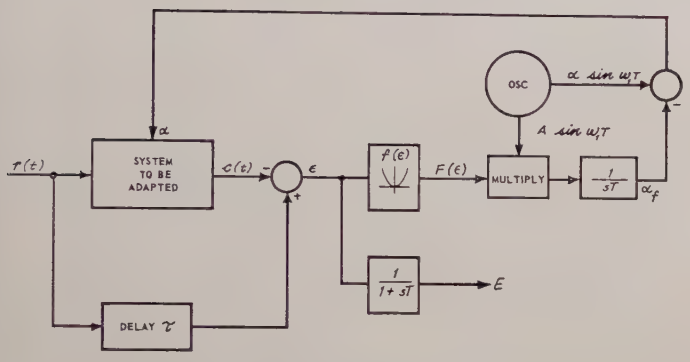


Fig. 18—Analog system for a parameter-perturbation adaptive system.

where $\epsilon_0(t)$ is the error for $\alpha_1 = 0$, and m_1 is the part of the modulation coefficient, the sign and size of which depend upon the misadjustment of the parameter α_1 for the conditions under which the system is operating.

The use of an error-squaring device here gives an output

$$\epsilon^2(t) \cong \epsilon_0^2(t)[1 + 2\alpha_1 m_1 \sin \omega_1 t], \tag{45}$$

ignoring terms in $\alpha_1^2 m_1^2$. After multiplication by $A \sin \omega_1 t$ and integration,

$$\alpha_f = \frac{1}{T}\int_0^T \epsilon_0^2(\sigma)[\sin \omega_1 \sigma + \alpha_1 m_1 - \alpha_1 m_1 \cos 2\omega_1 \sigma]d\sigma. \tag{46}$$

The second term in brackets gives the desired output, and this part of the return signal in the adaptive loop, if the loop time-constant is T', may be approximated by

$$\alpha_f \cong \epsilon_0^2(t)\alpha_1 m_1 T'/T. \tag{47}$$

Thus the loop gain depends upon the mean square error ϵ_0^2, as well as the modulation term, m_1. Thus, any large changes in error amplitude must be countered by either some kind of gain control, or by a limiter just ahead of the integrator.

One advantage of the scheme of Fig. 18 is that it permits the output of the squarer to be either integrated, or filtered to yield E, an estimate of the mean square error.

The feedback signal is dependent for sign and for size, as well, if limiting is not used, upon the modulation m_1, and this, in turn, will depend in general upon the function F. In the case of Gaussian input and a linear system, however, we can use any symmetrical function for F and the system will arrive at the same final value, since adaptation can only reduce τ. One choice for F in the linear system-Gaussian input case is a logarithmic criterion, with a modification to avoid negative values. As shown in Fig. 19, this is

$$\begin{aligned} F(\epsilon) &= \log\{1 + k|\epsilon|\} \\ &= \log\{1 + k|\epsilon_0|[1 + m_1\alpha_1 \sin \omega_1 t]\} \\ &\doteq \log[1 + k|\epsilon_0|] + \frac{k|\epsilon_0|}{1 + k|\epsilon_0|} m_1\alpha_1 \sin \omega_1 t. \end{aligned} \tag{48}$$

This function tends to give instantaneous limiting, in effect, because the term in $\sin \omega_1 t$ is independent of error amplitude, if $k|\epsilon_0| \gg 1$.

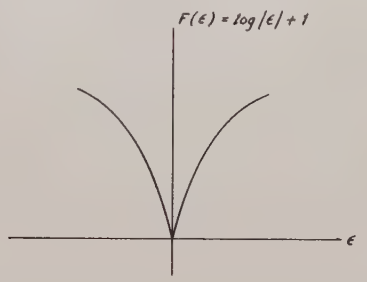


Fig. 19—Approximation to a logarithmic criterion curve.

However, this scheme makes desirable, for example, a separate squarer and filter, if it is desirable that the mean square error be conveniently monitored; and this becomes essential if the input is not Gaussian, or if the system is nonlinear. Unfortunately, it is in just such cases that the shape of the function F becomes important, because the modulation coefficient m_1 will be determined by the interrelationship between the probability density of the error and the function F. Here we are faced with a choice between shaping F to give instantaneous limiting, and shaping it to give some weighted minimum of the moments of ϵ. The latter may be desirable, and then other control of the gain of the adaptive loop may be desirable.

Investigations to extend both types of simulation described above are currently underway. Among the topics to be studied in both cases is the effect of the choice of the performance measure in the adaptive system. It is expected that further application of probabilistic concepts will be studied, especially in regard to probability-density-shaping techniques.

IV. Conclusion

Trends in the applications of performance measures have indicated the increased use of sophisticated mathematical concepts. Of particular importance is the use of statistical methods, in order that intuitive concepts can be aided or replaced by concepts that are more

mathematical in nature. Dynamic programming techniques and the application of correlation and translation functions are other methods that are becoming more important.

The application of performance measures to adaptive systems increases the difficulty of choosing a performance measure. We must consciously think about measuring the performance of the chosen performance measure, and perhaps even consider mechanizing this second-order measure. For example, if an adaptive control system with random inputs is subject to step-changes in a parameter, we may wish to have the adaptively controlled parameters assume their new values rapidly and without overshoot. An ITAE measure might be applied to the ensemble average of $f(\epsilon)$ as a function of time after the step change, to assist in the choice of the function $f(\epsilon)$ used in the adaptive loops. Such ideas and their mechanization may be extended to any practical limit.

All of this may seem discouraging to anyone seeking definite answers on performance measures. It seems apparent that there is no one magic formula, but that rather we are faced with a problem which is basically unanswerable. Nevertheless, engineers *must* choose performance measures, striving on the one hand to keep intuitive procedures from smothering the growth of more solidly based principles of measurement, and yet attempting to keep mathematical complexity from hiding simple meanings.

V. Acknowledgment

Assistance from members of the Electronics Department of Cornell Aeronautical Laboratory is gratefully acknowledged.

References

[1] A. C. Hall, "The Analysis and Synthesis of Linear Servomechanisms," The Technology Press, Mass. Inst. Tech., Cambridge, p. 19; 1943.
[2] I. Obradovic, "The deviation area in quick-acting regulation," *Archiv für Electrotechnik*, vol. 36, pp. 382–390; June, 1942. (In German.)
[3] N. Weiner, "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series," John Wiley and Sons, Inc., New York, N. Y.; 1948.
[4] A. R. Bergen, "A non-mean-square-error criterion for the synthesis of optimum finite-memory sampled data filters," 1957 IRE National Convention Record, pt. 2, pp. 26–32.
[5] T. R. Benedict and M. M. Sondhi, "On a property of Wiener filters," Proc. IRE, vol. 45, pp. 1021–1022; July, 1957.
[6] S. Sherman, "Non-mean-square error criteria," IRE Trans. on Information Theory, vol. IT-4, pp. 125–126; September, 1958.
[7] R. C. Oldenbourg and H. Sartorious, "The Dynamics of Automatic Controls," book published by the ASME, p. 66; 1948.
[8] C. Mack, "Calculation of the optimum parameters for a following system," *Phil. Mag.*, vol. 40, pp. 922–928; September, 1949.
[9] T. M. Stout, "A note on control area," *J. Appl. Phys.*, vol. 21, pp. 1129–1131; November, 1950.
[10] J. Laning and R. H. Battin, "Random Processes in Automatic Control," McGraw-Hill Book Co., Inc., New York, N. Y.; 1956.
[11] P. T. Nims, "Some design criteria for automatic controls", *Trans. AIEE*, vol. 70, pt. 1, p. 606–611; 1951.
[12] F. C. Fickeisen and T. M. Stout, "Analog methods for optimum servomechanism design," *Trans. AIEE*, vol. 71, pt. 2, pp. 244–250; November, 1952.
[13] R. R. Caldwell and V. C. Rideout, "A differential-analyzer study of certain nonlinearly damped servomechanisms," *Trans. AIEE*, vol. 72, pt. 2, pp. 165–169; July, 1953.
[14] D. Graham and R. C. Lathrop, "The synthesis of 'Optimum' transient response: criteria and standard forms," *Trans. AIEE*, vol. 72, pt. 2, pp. 278–288; November, 1953.
[15] J. H. Wescott, "The minimum-moment-of-error-squared criterion: a new performance criterion for servos," *Proc. IEE*, vol. 101, pp. 471–480; October, 1954.
[16] A. L. Whiteley, "The theory of servomechanisms, with particular reference to stabilization," *Proc. IEE*, vol. 93, pp. 353–372; August, 1946.
[17] H. M. James, N. B. Nichols, and R. S. Phillips, "Theory of Servomechanisms," McGraw-Hill Book Co., Inc., New York, N. Y., 1947.
[18] J. G. Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill Book Co., Inc., New York, N. Y., pp. 413–414; 1955.
[19] J. E. Gibson, "How to specify the performance of closed loop systems," *Control Engrg.*, vol. 3, pp. 122–129; September, 1956.
[20] W. C. Schultz and V. C. Rideout, "The selection and use of servo performance criteria," *Trans. AIEE*, vol. 76, pt. 2, pp. 383–388; 1957.
[21] W. C. Schultz, "Choice of Measures and Criteria for System Synthesis and Analysis," Ph.D. dissertation, University of Wisconsin, Madison; 1958.
[22] P. R. Aigrain and E. M. Williams, "Design of optimum transient response amplifiers, Proc. IRE, vol. 37, pp. 873–879; August, 1949.
[23] Y. W. Lee and J. B. Wiesner, "Correlation functions and communications applications," *Electronics*, vol. 23, pp. 86–92; June, 1950.
[24] M. G. Spooner and V. C. Rideout, "Correlation Studies of Linear and Nonlinear Systems," paper presented at the Natl. Electronics Conf., Chicago, Ill.; October, 1956.
[25] G. C. Newton, Jr., L. A. Gould, and J. F. Kaiser, "Analytical Design of Linear Feedback Controls," John Wiley and Sons, Inc., New York, N. Y., pp. 110–113; 1957.
[26] V. C. Rideout, "Some Applications of a High-Speed Analog Correlator," presented at the Natl. Simulation Conf., Dallas, Tex.; October 23–25, 1958.
[27] L. A. Zadeh and J. R. Ragazzini, "An extension of Wiener's theory of predictions," *J. Appl. Phys.*, vol. 21, pp. 645–655; July, 1950.
[28] I. Kanter, "Some new results for the prediction of derivatives of polynomial signals in additive stationary noise," 1959 IRE WESCON Convention Record, pt. 4, pp. 87–91.
[29] B. Widrow, "Adaptive sampled-data systems—a statistical theory of adaptation," 1959 WESCON Convention Record, pt. 4, pp. 74–85.
[30] R. C. Booton, Jr., "Nonlinear control systems with random inputs," IRE Trans. on Circuit Theory, vol. CT-1, pp. 9–18; March, 1954.
[31] H. S. Hsieh and C. T. Leondes, "On the optimum synthesis of sampled data multipole filters with random and nonrandom inputs," 1960 IRE International Convention Record, pt. 4, pp. 37–52.
[32] H. R. Leland, "Response of Certain Nonlinear Systems to Random Inputs," Ph.D. dissertation, University of Wisconsin, Madison; 1958.
[33] G. S. Axelby, "Random noise with bias signals in nonlinear devices," 1959 WESCON Convention Record, pt. 4, pp. 171–185.
[34] W. M. Kaufman, "The Use of Generalized Error Criteria in the Statistical Evaluation of System Performance," Elec. Engrg. Dept., Carnegie Inst. Tech., Pittsburgh, Pa., Progress Rept., ONR Contract N7onr 30306.
[35] J. Zaborszky and J. W. Diesel, "Probabilistic error as a measure of control-system performance," *Trans. AIEE*, vol. 78, pt. 2; July, 1959.
[36] J. Zaborszky and J. W. Diesel, "A statistically averaged error criterion for feedback-system synthesis," *J. Aero/Space Sci.*, pp. 128–134; February, 1960.
[37] J. B. Reswick, "Determine system dynamics—without upset," *Control Engrg.*, vol. 2, pp. 50–57; June, 1955.
[38] G. W. Anderson, J. A. Aseltine, A. R. Marcini, and C. W. Sarture, "A self-adjusting system for optimum dynamic performance," 1958 IRE National Convention Record, pt. 4, pp. 182–190.
[39] Proc. Symp. on Self-Adaptive Control Processes, Wright Air Dev. Ctr., Wright-Patterson AFB, Dayton, Ohio, Tech. Rept. 59–49, ASTIA Doc. No. AD-209389; March, 1959.
[40] Y. W. Lee, "Application of Statistical Methods to Communications Problems," Mass. Inst. Tech., Lexington, Res. Lab. for Electronics, Tech. Rept. 181; September, 1950.
[41] H. M. Paynter, "On an analogy between stochastic processes and monotone dynamic systems," in "Regelungstechnik Moderne Theorien und Ihre Verwendborheit," Oldenbourg, Munich, Germany; 1957.
[42] T. P. Goodman and R. H. Hillsley, "Continuous measurement of characteristics of systems with random inputs: a step toward self-optimizing control," *Trans. ASME*, vol. 80, pp. 1839–1848; November, 1958.

[43] G. J. Murphy and N. T. Bold, "Optimization based on a square-error criterion with an arbitrary weighting function," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 24–30; January, 1960.
[44] R. Bellman, "Dynamic Programming," Princeton University Press, Princeton, N. J.; 1957.
[45] R. Bellman and R. Kalaba, "On adaptive control processes," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-4, pp. 1–9; November, 1959.
[46] R. Bellman and R. Kalaba, "Dynamic programming and adaptive processes: mathematical foundation," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 5–10; January, 1960.
[47] M. Freimer, "A dynamic programming approach to adaptive control processes," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-4, pp. 10–15; November, 1959.
[48] C. W. Merriam, III, "Use of a mathematical error criterion in the design of adaptive control systems," *Trans. AIEE*, vol. 79, pp. 506–512; January, 1960.
[49] C. W. Merriam III, "A class of optimum control systems," *J. Franklin Inst.*, vol. 267, pp. 267–281; April, 1959.
[50] R. F. Kalman and R. W. Koepcke, "Optimal synthesis of linear sampling systems using generalized performance indexes," *Trans. ASME*, vol. 80, pp. 1820–1826; November, 1958.
[51] L. A. Zadeh, "What is optimal?" IRE TRANS. ON INFORMATION THEORY, vol. I-4, p. 3; March, 1958.
[52] R. F. Baum, "The correlation function of smoothly limited Gaussian noise," IRE TRANS. ON INFORMATION THEORY, vol. IT-3, pp. 193–197; September, 1957.
[53] W. L. Still, "Separate signal from noise with probability filters," *Control Engrg.*, vol. 7, pp. 147–151; March, 1960.
[54] T. R. Benedict, H. R. Leland, W. C. Schultz and M. G. Spooner, "A Study of an Adaptive Control System Using a Digital Simulation," presented at the Natl. Specialists Meeting on Guidance of Aerospace Vehicles, Inst. Aeronautical Sci., Boston, Mass.; May 23–27, 1960.
[55] R. J. McGrath and V. C. Rideout, "A simulator study of a two-parameter adaptive system," this issue, pp. 35–42.
[56] H. J. Milsum, "Statistical optimization of regulators employing a binary error criterion," *Trans. ASME*, Paper No. 58-A-71.
[57] H. von H. Witsenhausen, "Anwendung von analogierechenmaschinen auf die optimierung von unstetigen regelkreisen mit statistisch schwankenden eigangsgrossen," paper presented in Duesseldorf, Germany; August, 1957.
[58] A. I. Topitsyn, "An integral estimate for selecting the optimum of an automatic control system with a given overshoot," *Automation and Remote Control*, vol. 20; April 1959.
[59] E. W. James and A. S. Boksenbom, "How to establish the control problem for an on-line computer," *Control Engrg.*, vol. 4, pp. 148–159; September, 1957.
[60] M. G. Spooner, "The Use of Correlation Techniques in the Study of Linear and Nonlinear Servomechanisms," Ph.D. dissertation, University of Wisconsin, Madison; 1956.
[61] H. H. Rosenbrock, "Integral-of-Error-Squared Criterion for Servomechanism," *Proc. IEE*, vol. 102, pp. 602–607; September, 1955.
[62] T. R. Benedict and V. C. Rideout, "Error determination for optimum predicting filters," *Proc. NEC*, vol. 13, pp. 875–887; 1957.
[63] W. C. Schultz and V. C. Rideout, "A general criterion for servo performance," *Proc. NEC*, vol. 13, pp. 549–560; 1957.
[64] R. N. Clark, "Integral of error squared as a performance index for automatic control systems," *Trans AIEE*, Paper 60-1016, vol. 79; August, 1960.
[65] J. Zaborszky and J. W. Diesel, "Design of continuous linear control system for minimum probabilistic error," *Trans AIEE (Applications and Industry)*, vol. 79, pp. 44–54; May, 1960.
[66] J. Zaborszky and J. W. Diesel, "Design of sampled-data control systems for minimum probabilistic error," *Trans. AIEE (Applications and Industry)*, vol 79, pp. 54–65; May, 1960.
[67] W. E. VanderVelde, "Make statistical studies on analog simulators," *Control Engrg.*, vol. 7, pp. 127–130; June, 1960.
[68] J. E. Gibson, "Making sense out of the adaptive principle," *Control Engrg*, vol. 7, pp. 113–119; August, 1960.

A Simulator Study of a Two-Parameter Adaptive System*

R. J. McGRATH†, MEMBER, IRE, AND V. C. RIDEOUT‡, FELLOW, IRE

Summary—**The use of sinusoidal parameter perturbation applied to a feedback control system to obtain an adaptive scheme which optimizes the system for changes in inputs and/or system parameters is discussed. It is shown that if a parameter perturbation signal is cross-correlated with the system error squared, the correlator output can be used to adjust the parameter to minimize the mean-square error. Other error measures may also be used. Two or more parameters may be simultaneously adjusted if they are perturbed at different frequencies, and each provided with an independent adaptive loop. A computer simulation of a third-order system having two adjustable parameters was examined for a variety of inputs including random signals. It is shown that the scheme minimizes the mean-square error in all cases.**

* Received by the PGAC, May 25, 1960; revised manuscript received, October 21, 1960. This work was aided by support from the Wisconsin Alumni Res. Foundation, the Wisconsin Engrg. Experiment Sta., the Natl. Science Foundation, and the Office of Ordnance Res., U. S. Army, Contract No. DA-11-022-ORD-2059.

† Aerospace Corp., Los Angeles, Calif. Formerly with Electrical Engrg. Dept., University of Wisconsin, Madison.
‡ Electrical Engrg. Dept. and Math. Res. Center of the U. S. Army, University of Wisconsin, Madison.

I. INTRODUCTION

SYSTEMS such as feedback control devices may be designed to give responses which are optimum in some sense (*e.g.*, minimum mean-square error) for a variety of different input signals, each occurring with known probability. Better results will be obtained if those parameters in the system which can be varied are so adjusted as to optimize the output for each class of input signal. This may be done by preprogramming if the class of signals to be expected at any given time is known. If this information is not available then a "signal-adaptive" system may be set up, which analyzes or "identifies" the input signal and automatically optimizes the system as the input signal statistics change. Of course there may be more than one input signal, and some of these may be disturbances rather than commands.

Another problem results if any of the parameters of

the system change with time. If this change cannot be countered by preprogrammed changes in the system, then a control loop may be desirable to change the adjustable parameters of the system so that the output is optimized according to some chosen criterion. Such a system may be said to be "system-adaptive."

The study described in this paper is one which uses a scheme of parameter perturbation and correlation detection in feedback loops, resulting in a self-adaptive feature. Quasi-stationary random inputs, whose statistics change slowly compared to periods of at least their most important spectral content, and similarly slow random variations in system parameter values were assumed, although step changes were convenient for many simulator studies. The system is applicable to either analog or digital methods of control.

This scheme for adaptation was first mentioned by Draper and Li [1]–[3]. The studies of the optimization of process control systems carried out in early studies by Box [4] are related to this work, except that, in the systems he describes, a human is used to close the adaptive loop. Widrow [5] has described an adaptive sampled-data system using step changes. Bibliographies by Aseltine, *et al.* [2], and Stormer [6] cover a number of types of adaptive systems.

II. System Description

The adaptive system, as shown for the single parameter case in Fig. 1, consists of a parameter controller, error function device, filter, integrator, and multiplier, all connected to the system in what might be called a parametric feedback loop. A small sinusoidal perturbation is applied to the parameter α, which is to be controlled. The frequency ω_1 of the perturbation may be intermediate between the frequencies in the error spectrum (ignoring its nonstationarity) and the low frequencies in system parameter disturbances or in signal statistics changes.

The error function device of Fig. 1 gives some nonnegative zero-memory function of the error. It might, for example, be the error squared,

$$F(\epsilon) = \epsilon^2. \tag{1}$$

We are concerned with the first moment of ϵ^2, which may be estimated by time integration. Ignoring, initially, the uncertainty in such a measurement (due to finite time of integration, and low-frequency nonstationarity)

$$E = \overline{F(\epsilon)} \approx \frac{1}{T}\int_0^T F(\epsilon)dt, \qquad T \text{ large.} \tag{2}$$

This quantity, which is a measure of error, is a metric and may be regarded as a function of the input statistics and of the variable system parameters. If only one parameter α is varied, and $E(\alpha)$ tends to be a parabolic function of α as shown in Fig. 2, then it can be

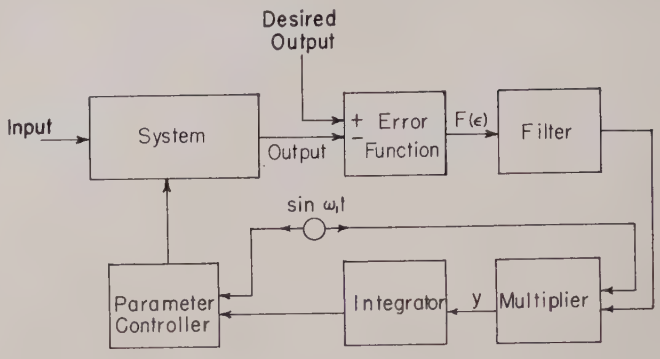


Fig. 1—Parameter perturbation adaptive system in block form.

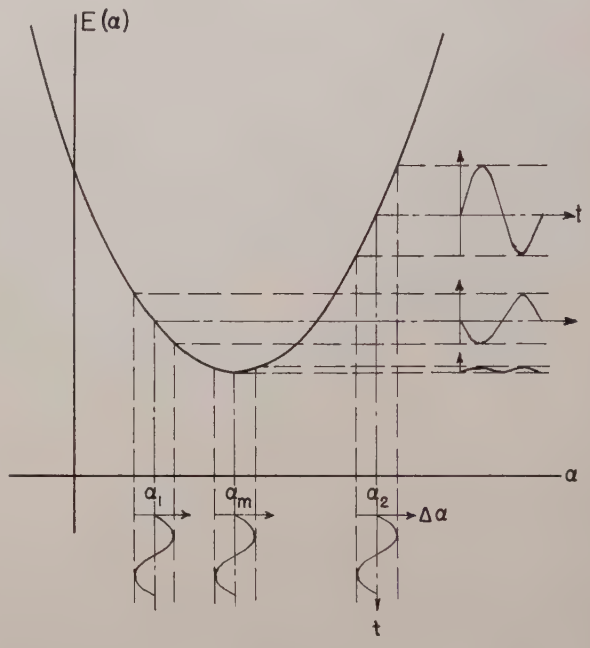
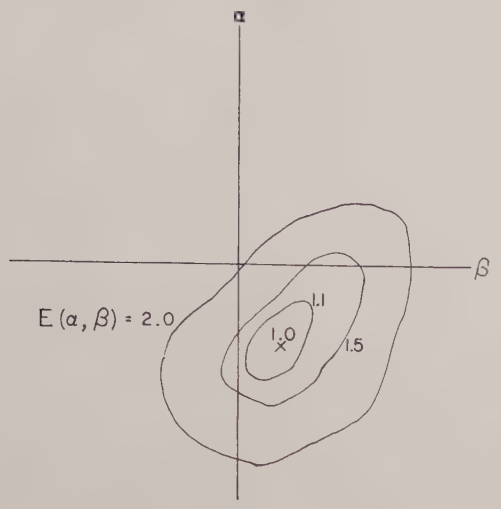


Fig. 2—Average of an error function $E(\alpha)$, as one parameter α is varied.

seen that in the stationary case the output at frequency ω_1 will tend to be of one phase if α is below its optimum value, and of opposite phase if it is above (cases 1 and 2 in Fig. 2). If α is at the value α_m which minimizes $E(\alpha)$ there will be no output at frequency ω_1.

The correlation detector, or phase-sensitive detector, made up of a multiplier and integrator combines the perturbation signal and the filtered error measure to give a signal which will cause α to approach α_m. Of course changes in input statistics, or changes in other parameters, may change the position and shape of the curve of $E(\alpha)$ in Fig. 2, but the system will continually try to seek that value of α_m corresponding to the instantaneous minimum of this curve.

If more than one parameter of the system is to be controlled, then a separate loop must be set up for each parameter using different perturbation frequencies. The same error measure may be used for each loop if this is appropriate, but other elements must be duplicated for each parameter. It is not essential that the error be an



F g. 3—Contours of equal error-function average $E(\alpha, \beta)$ for the two-parameter case.

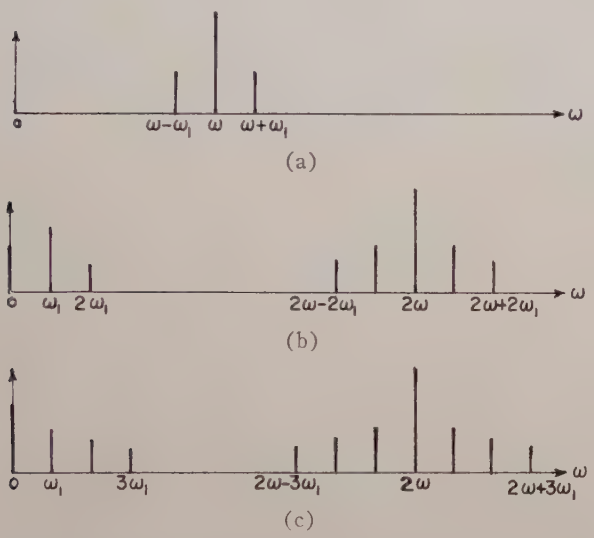


Fig. 4—Spectra in a parameter-perturbation adaptive system with sinusoidal input. (a) Spectrum of error. (b) Spectrum of error squared. (c) Spectrum of multiplier output, $y = \epsilon \sin \omega_1 t$.

orthogonal function of the parameters. Fig. 3 shows a realistic set of contours for a two-parameter case.

It is desirable that there be a single minimum in the error measure, viewed as a function of the adjustable parameters. This is not the general case, although a single minimum is often likely over the important ranges of parameter values even for large numbers of parameters [7]. If multiple minima are troublesome, an auxiliary model technique could be used to find an absolute minimum.

III. Analysis of the Adaptive Loop—Sinusoidal Input Case

Consider first the case in which a single adaptive loop is to be added to a linear system to minimize the average of some function of the error by adjustment of a parameter α. If the adaptive loop (Fig. 1) is opened at the integrator output, and α has the initial value α_0, then with a perturbation of some small amplitude α_1,

$$\alpha = \alpha_0 + \alpha_1 \sin \omega_1 t. \tag{3}$$

The instantaneous error may be assumed to be modulated by this perturbation,

$$\epsilon(t) = [1 + \alpha_1 m \sin \omega_1 t]\epsilon_0(t) \tag{4}$$

where $\epsilon_0(t)$ is the error for $\alpha = \alpha_0$ in the absence of modulation, and the degree of modulation, $\alpha_1 m$, is small.

For the sinusoidal input case,

$$r(t) = R \sin \omega t, \tag{5}$$

where ω is, typically, a higher frequency than the perturbation frequency ω_1. Now the error with no modulation will be of the form

$$\epsilon_0(t) = k_0 R \sin (\omega t + \phi_0) \tag{6}$$

because of linearity. We will ignore ϕ_0, or assume it zero, since it is unimportant in the analysis which follows. The modulated error is

$$\epsilon(t) = k_0 R(1 + \alpha_1 m_1 \sin \omega_1 t) \sin \omega t. \tag{7}$$

Although it is interesting and important to consider many error criteria, we will take the way of greatest mathematical ease and consider first the mean-square error criterion. The square of (7), after multiplication by $A_1 \sin \omega_1 t$, gives the integrator input, y, (Fig. 1).

$$y = \frac{k_0^2 R^2 \alpha_1 m_1 A_1}{2} + \text{sinusoidal terms}, \quad \text{if } \omega_1 < 2\omega/3. \tag{8}$$

Fig. 4 shows the spectra for ϵ, ϵ^2, and y for the sinusoidal input case.

A long integration of y will tend to make all oscillatory terms in (8) insignificant in comparison with the constant first term.

Assume a parabolic form for the mean-square error,

$$\epsilon_0^2 = R^2[P + Q(\alpha_0 - \alpha_m)^2], \tag{9}$$

such that ϵ_0^2 has a minimum value $R^2 P$ at $\alpha_0 = \alpha_m$. It can be shown that the modulation index is now

$$m_1 = \frac{2Q(\alpha_0 - \alpha_m)}{k_0^2}. \tag{10}$$

Substitution of this value for the modulation coefficient m_1 into the nonsinusoidal first term in (8) yields

$$y_1 = R^2 \alpha_1 A_1 Q(\alpha_0 - \alpha_m). \tag{11}$$

Thus, for the assumed parabolic mean-square error curve, the dc part of the signal applied to the integrator is proportional to the difference, $(\alpha_0 - \alpha_m)$, between the actual value of the parameter α_0, and α_m, the value which makes $\overline{\epsilon_0^2}$ a minimum.

Also, it is to be noted that y_1 in (11) is proportional to the derivative of $\overline{\epsilon_0^2}$ with respect to α_0; from (9),

$$\frac{d\overline{\epsilon_0^2}}{d\alpha_0} = 2R^2 Q(\alpha_0 - \alpha_m). \tag{12}$$

Therefore,

$$y_0 = K\frac{d\overline{\epsilon_0^2}}{d\alpha_0}. \tag{13}$$

In general, for any error measure E and any number of parameters, the parabola which fits E in an α-plane near α_0 will give

$$y_0 = K\frac{\partial E(\alpha, \beta, \cdots)}{\partial \alpha_0}, \tag{14}$$

although K will tend to be a function of α_0 for nonparabolic curves.

Some information of importance regarding the adaptive loop may be deduced from (12). Of most importance is the presence of the term $(\alpha_0 - \alpha_m)$ indicating that the system plus squarer and correlating multiplier is equivalent to a subtraction circuit yielding the error in the parameter α with respect to its optimum value. It can be seen that y will have the same form if α_m is changed to a new value either because the frequency of the input or the value of some other parameter has been changed. In these cases Q may also change, thus changing the gain of the adaptive loop. A more serious change in gain appears if the input $r(t)$ changes level, for y varies as R^2 varies. This is undesirable because of possible instability. The open loop gain of the adaptive loop is approximated by

$$Y(p) = \frac{R^2\alpha_1 A_1 Q}{pT} \tag{15}$$

from (11) and Fig. 1, where T is the integrator time constant. This approximate form indicates stability for all values of R, but neglected poles may lead to instability as R increases. This difficulty may be overcome by:

1) use of automatic gain control in the adaptive loop,
2) use of a limiter ahead of the integrator,
3) use of a criterion [12] other than mean square which serves as an instantaneous limiter.

IV. Analysis of the Adaptive Loop —Random Inputs

Suppose that the system which is to be made self-adaptive is a linear servomechanism with a random input such that the error spectrum tends to be concentrated at the higher frequencies contained in the input spectrum. In such cases it was felt that the perturbation frequencies should be as high as possible, but below the main part of the error spectrum.

It might be thought that in such a system, adaptive loop response time would be primarily governed by the length of a perturbation period. It is easy to show that this is not the only consideration for the random input case by considering the effect of a square-wave perturbation of period τ, for the Gaussian input case. If the loop is opened at the integrator output, the normalized variance of this output at the end of the first half-cycle may be approximated [9] by

$$\sigma^2 \approx 1/F\tau, \tag{16}$$

where F is the bandwidth of the error $\epsilon(t)$ and is assumed to be narrow relative to its midband. If this variance is not sufficiently small, then since F is fixed by the system we must *increase* τ, or use, in effect, the average of many cycles of the square-wave perturbation. The latter is preferable, because τ should be kept small enough to reduce variations in the parameter due to perturbation and to follow more quickly changes in input statistics or parameters.

Thus the only means which can be used to reduce the variance of the integrator output is an increase in the number of perturbation cycles used, or, in other words, an increase in the averaging time in the integrator. This is equivalent to a decrease in loop gain, which will give slower closed-loop response. Some optimum must therefore be sought which is a compromise between slow but steady response and faster but more noisy or uncertain response of the adaptive loop. Such compromises, because of the complexity of the actual systems, may best be sought by analog simulation methods (see Section V below).

As we go from the square-wave perturbation assumed above to the sinusoidal perturbation actually used in these studies, the modulated error may be assumed to be of the form given in (4). Its square will be

$$\epsilon^2(t) = \epsilon_0^2(t)[1 + 2\alpha_1 m_1 \sin \omega_1 t + \alpha_1^2 m_1^2 \sin^2 \omega_1 t]. \tag{17}$$

The integrator output will be, if $z(0) = 0$,

$$z(t) = \frac{A_1}{T}\int_0^t \epsilon_0^2(\rho)[\sin \omega_1\rho + \alpha_1 m_1 - \alpha_1 m_1 \cos 2\omega_1\rho]d\rho, \tag{18}$$

where terms in $(\alpha_1 m_1)^2$ have been neglected as being very small compared to other terms. Let us further assume that integration proceeds up to a time $t = T'$, where T' is the effective time constant of the adaptive loop. Since m_1 and $\overline{\epsilon_0^2}(t)$ cannot be changed greatly by adaptive loop transmission in time T', an open loop will be assumed, initially, and we have

$$z(T') \approx \frac{A_1}{T}\int_0^{T'} \epsilon_0^2(\rho)[\sin \omega_1\rho + \alpha_1 m_1]d\rho, \tag{19}$$

where the term of frequency $2\omega_1$ is small because $T' \gg 1/\omega_1$, in general, and $\alpha_1 m_1$ is small. An ensemble average of $z(T')$ is

$$\langle z(T') \rangle = \frac{A_1}{T}\int_0^{T'} \langle \epsilon_0^2 \rangle [\sin \omega\rho + \alpha_1 m_1]d\rho$$

$$\approx \frac{A_1}{T}\overline{\epsilon_0^2}\left[\frac{1 - \cos \omega_1 T'}{\omega_1} + \alpha_1 m_1 T'\right], \tag{20}$$

where a short-time estimate of ϵ_0^2 replaces the ensemble average. If $\omega_1 T' \gg 1$, then

$$\langle z(T') \rangle = A_1 \overline{\epsilon_0^2} \alpha_1 m_1 T'/T. \tag{21}$$

This is of the same form as (8), times T', and indicates that the adaptive loop will tend to be effective in the random input as well as the sinusoidal input case.

It is difficult to carry this analysis further in a meaningful fashion, although an attempt has been made to calculate the variance of $z(T')$ in this open-loop case [10]. In order to do so, however, it is in turn necessary to know the autocorrelation of the error. Such an analysis indicates that a filter following the squarer, and centered at ω_1, will have no first-order effect on the variance of $z(T')$ but may give some slight reduction of $z(T')$ due to second-order effects. This result is borne out by experiment, and, since filtering is desirable in any case to protect the multiplier from overload, it has been made a standard part of each adaptive loop. Some care must be exercised that the filter is not so narrow that added poles in the loop can result in instability.

The adaptive loop output, ignoring random variations, may be obtained from (21) by substituting the operator $1/p$ for T':

$$z = \alpha_0 = \frac{A_1 \overline{\epsilon_0^2} \alpha_1 m_1}{pT}, \tag{22}$$

where as before (10) a subtraction is inherent in m_1, *i.e.*,

$$m_1 = K_1(\alpha_0 - \alpha_m), \tag{23}$$

$$G = \frac{\alpha_0}{\alpha_0 - \alpha_m} = \frac{A_1 \overline{\epsilon_0^2} \alpha_1 K_1}{pT}. \tag{24}$$

Thus, ignoring poles resulting from the addition of the filter, and from imperfections in multiplier and integrator, the loop-gain characteristic is such as to be always stable. However, it is proportional to $\overline{\epsilon_0^2}$.

A variation in mean-square error will necessarily occur as the error is reduced by adaptive control. Of more importance is the variation in $\overline{\epsilon_0^2}$ as input signal or disturbance level changes. Some improvement can be gained by choice of an error function other than mean square [12]. The mean absolute value criterion, for example, is less sensitive to amplitude changes than the mean square.

An effective scheme for minimization of the effects of change in input signal amplitude on $\overline{\epsilon_0^2}(t)$, and thus on loop gain, has been to include a high-gain limiter at the output of the multiplier. This gives, in effect, a relay regulator, and furthermore it is one in which gain is easily controlled by setting the limiter amplitude.

It will be noted in Fig. 7 that both the filter and the limiter discussed in this section have been included in each adaptive loop of the system studied by analog computer methods.

One problem remaining in this system is to guard against misadjustment of parameters by the adaptive loop when the error signal amplitude is so low as to be unreliable for adaptive loop control. One possibility in this case is to introduce a dead-space in the high-gain limiters, so that the integrators will hold the value of the parameter reached before a reduction of error amplitude to some predetermined value. Switching and restoration of parameters to some chosen zero-signal value is also possible when the error signal sinks below some threshold.

V. Experimental Results

The parameter-perturbation adaptive system was studied with the aid of an analog computer simulation of a third-order system,

$$\frac{C(p)}{R(p)} = \frac{K(p\beta + 1)}{(pB + 1)[p^2 + p(A + K\alpha) + K]}, \tag{25}$$

shown in block form in Fig. 5. Variations of A, B, K and of the spectrum $\Phi_{rr}(j\omega)$ were assumed to be possible, with compensation to be made by self-adaptive adjustment of the controllable parameters α and β. The input $r(t)$ was assumed to be the desired output.

The computer used was the Wisconsin-Philbrick, built for the most part at the University of Wisconsin, using Philbrick plug-in amplifiers. It operates normally on a millisecond time base, making it a practical matter to process random data. Fifty-five operational amplifiers, seven multipliers, and a wide variety of nonlinear elements are available.

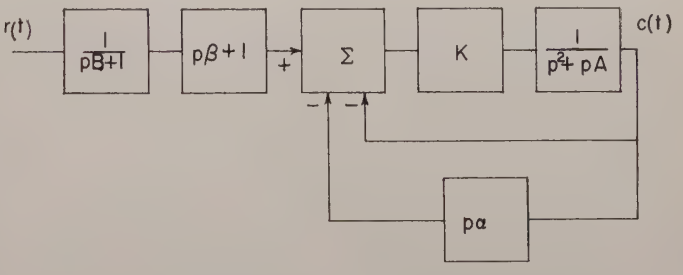


Fig. 5—Block diagram of third-order system used in simulation.

Fig. 6(a) shows the spectrum of ϵ^2 when the input $r(t)$ was a 200-cps sinusoid but the third-order system was adjusted for optimum square-wave response. Fig. 6(b) shows the effect of varying α at 47 cps, and Fig. 6(c) the effect of varying β at 31 cps. Fig. 6(d) shows the effect of varying both α and β. Note that components at the perturbing frequencies appear in ϵ^2. Tests were made with other settings for α and β with similar results.

Fig. 7 is a block diagram of the complete system with two adaptive loops. Electronic multipliers were used to change the parameter values, as the simulator frequencies were too high for a mechanical servo to be used to

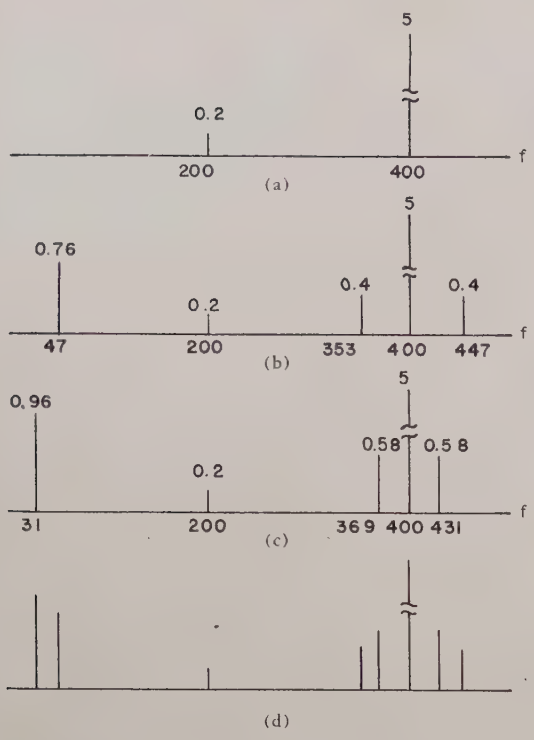


Fig. 6—Error-squared spectrum of third-order system with 200-cps sinusoidal input. (a) No perturbation. (b) α perturbed at 47 cps. (c) β perturbed at 31 cps. (d) α and β perturbed; amplitude unchanged.

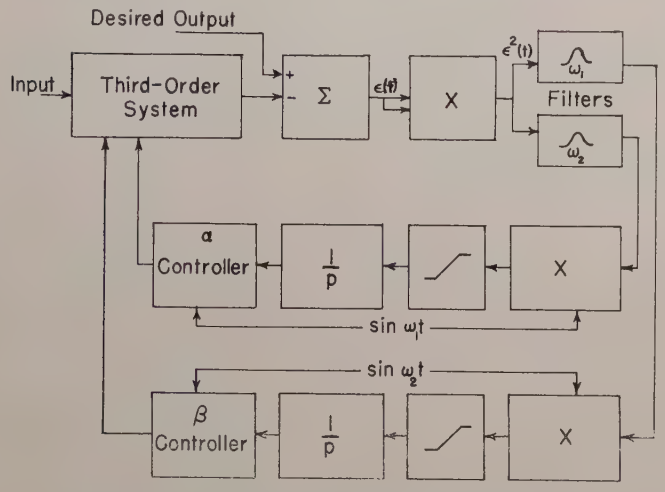


Fig. 7—Block diagram of parameter-perturbation scheme designed for adaptive control of two parameters.

control a pot setting. The bias from the parameter controller on each input channel then determined the parameter value. Again the input was regarded as desired output. With a random input the system would not work without the band-pass filters discussed in the preceding section. It should again be noted that these filters are essential to prevent overload of the multipliers. The output of the correlating multiplier in each loop was amplitude limited. The limiter was found to be desirable for random inputs, not only to counteract changes in gain which would otherwise result as error amplitude changes (Section IV), but also to reduce the effects of sudden large "false correlations" which otherwise would not only cause wide excursions in the parameters, but could actually cause "statistical" instability in the adaptive loop. For the adaptive loops with limiters included, the effective loop gain can easily be controlled by adjustment of the limiting level.

A 200 cps square wave was first used as the system input. Fig. 8 shows contours of equal mean-square error vs settings of α and β, the controllable parameters. In this figure the center of the grid is the point of minimum MSE. A change of one grid division on the α axis corresponds to a 25 per cent change in the parameter value. On the β axis one grid division corresponds to a 16 per cent change. Also shown in Fig. 8 are a series of dots culminating in a solid line. These curves are the paths the parameters take when the adaptive loop is operating, causing the parameters to move toward the optimum adjustment at the center of the grid. Initially the parameters were misadjusted, then the adaptive loop was turned on and they moved to their optimum position. The dots occur at one-tenth second intervals. (With the periodic input the adjustment speed can be made much faster than shown but the system was set up with loop gains which were suitable for random inputs.)

Fig. 8 shows that the action of the two adaptive loops is such as simultaneously to change α and β so as to follow a path of steepest descent on the MSE surface. Comparison of the adaptive loop control with human control of the parameter values showed substantial advantages in favor of the former mainly because of the difficulty that the human finds in minimization of a function which requires simultaneous variation of nonorthogonal or interacting parameters.

Fig. 9 shows parameter response vs time for the case discussed above. In Fig. 9(a) the parameter α is periodically disturbed by a square wave. The first quick recovery of α corresponds to the steep descent to the valley of the MSE surface, and the following slow recovery to the descent along the valley floor to its deepest point. Meanwhile, β shows excursions which result from the interactions between these nonorthogonalized parameters. (If the parameters were orthogonal, β would not change as α is periodically off-set.)

Fig. 9(b) shows the effect of periodically disturbing β with a square wave.

The MSE surfaces for two random inputs were next determined. White Gaussian noise was passed through a shaping filter and applied as the system input. Fig. 10(a) is for a shaping filter with a transfer function

$$\frac{e_0}{e_i} = \frac{1}{10^{-6}p^2 + 0.05 \times 10^{-3}p + 4}, \tag{26}$$

while the contours of Fig. 10(b) are for the filter

$$\frac{e_0}{e_i} = \frac{1}{(10^{-3}p + 2)^2}. \tag{27}$$

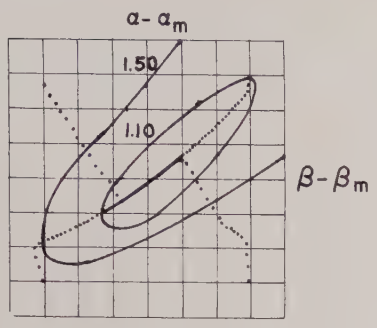


Fig. 8—Contours of equal MSE for square wave input. Also shown is the parameter adaptation towards the center of the grid (minimum MSE) from four different initial misadjustments.

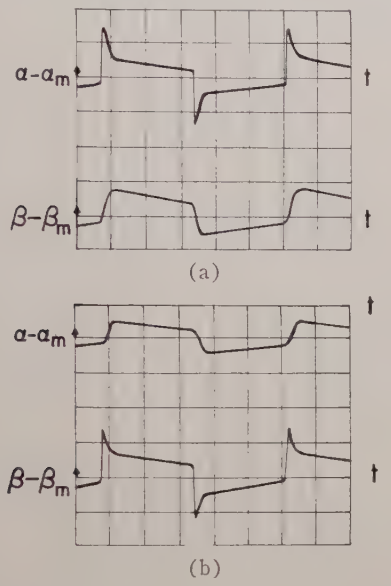


Fig. 9—Parameter response to step misadjustments. (a) Step forced on α. (b) Step forced on β.

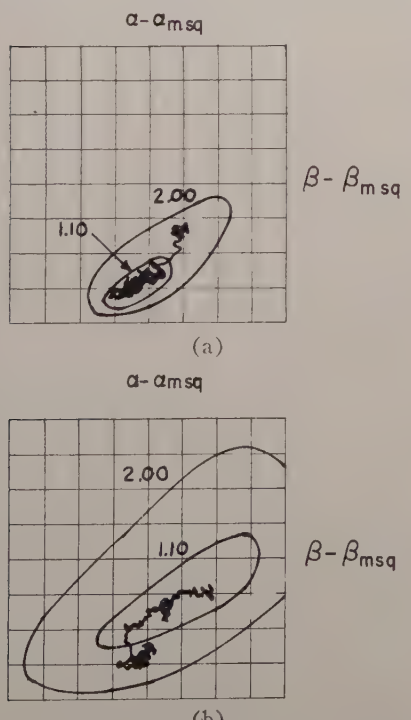


Fig. 10—Contours of equal MSE for two random signals. Also shown is the parameter adaptation as the inputs are changed from one to the other.

Fig. 10(a) also shows a trace of the parameters adjusting to a change in the input signal. In Fig. 10(a) the starting position is at the minimum for the filter of (27) and the input is changed to that of (26). Unlike the case of the square wave input, the adjustment does not move the point of operation directly to a minimum but "wanders" towards it. In Fig. 10(b) the filter sequence is reversed. These curves are continuous traces of the parameters' motion over a period of 30 seconds.

Fig. 11 shows the parameters adjusting to a change in input signal from the 200-cps square wave to the random output of the filter of (26). In this trace the dots occur at one-tenth second intervals. Thus with inputs in the 200–300-cps range it took about one second to reduce the MSE by a factor of four. If the adaptive loop gain is raised considerably, instability in the adjustment is possible. Therefore all factors affecting loop gain must be chosen with a compromise between speed and the amount of wander in the adjustment paths.

In an effort to check the operation of the adaptive system, the third-order system was reduced to second order with both β and B set to zero. With a square wave input α was allowed to adjust and its value measured. The sum of A and $K\alpha$ was checked against the computed value for minimum integral squared error (ISE) for a step [11]. Although the measured value was 3.2 per cent different from theoretical, this measured value results in less than a 0.1 per cent theoretical increase in ISE from the minimum. The ISE for a step input is a meaningful measure since the error essentially goes to zero in half of the square-wave period.

As a final test with the second-order system, a limiter was included in series with K in Fig. 5. Fig. 12(a) shows the square wave response with damping held at the optimum value for the case without limiting, while Fig. 12(b) shows the response when α is allowed to adapt. The value of α in this case was just twice that of the optimum value without limiting.

VI. Conclusion

It has been shown that a linear system can be made to be both signal-adaptive and system-adaptive, *i.e.*, self-optimizing with respect to nonstationary random inputs and random parameter changes, by use of a parameter-perturbation scheme. This scheme, which works well for two adjustable parameters, should be readily adapted to more than two, and orthogonalization of the parameters with respect to the error measure does not appear to be necessary. It also appears possible to extend this system to the adaptation of essentially nonlinear systems, as well as the simple saturating type of nonlinearity discussed in the preceding section.

Although the mean-square criterion was used in most of the experimental work described here, other criteria [12] may offer advantages, particularly where the error spectrum is non-Gaussian.

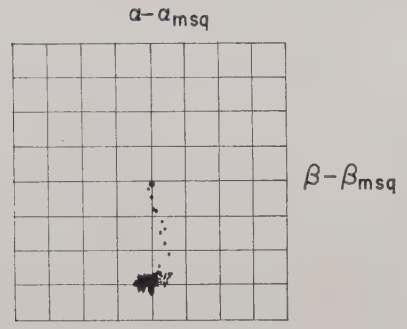


Fig. 11—Response of α and β to input change from square waves to the output of the filter of (27).

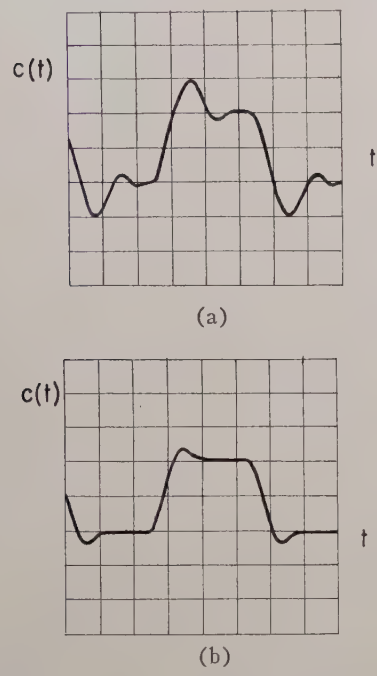


Fig. 12—Response of second-order system with saturation to a 200-cps square wave. (a) Adjusted for optimum response if there were no saturation. (b) Derivative feedback allowed to adapt.

The high degree of nonlinearity in this type of adaptive system makes analysis difficult and leads to much use of approximation. Analog simulation, particularly high speed or fast-time-scale analog simulation offers considerable advantages in this study, particularly where random inputs are concerned. This would be still more true where second-order adaptation, or adaptation of the adaptive loop itself, made averaging times still longer.

A general conclusion of importance in any application of this scheme is that it is not essential that the designer know in great detail the parameters or even the construction of the system to be adapted. Also, the methods described can be realized with discretely as well as continuously variable components.

The speed of response of any adaptive system is of key importance. In pure adaptation of the kind discussed in this paper (*i.e.*, where no preprogramming is possible), a certain amount of time lag is inevitable to permit "processing" of the information in the error function. In this system the parameters were well adapted in 1 to 2 seconds with inputs in the 200–300-cps range (Figs. 8 and 11). The adjustment time should be proportional to the reciprocal of the operating frequencies.

In more recent studies using slightly different techniques the adaptation time has been speeded up by factors of from 10 to 100. The authors hope to report on this in detail at a later date.

Bibliography

[1] C. S. Draper and Y. T. Li, "Principles of optimalizing control systems and an application to the internal combustion engine," ASME publication; September, 1951.
[2] J. A. Aseltine, A. R. Mancini and C. W. Sarture, "A survey of adaptive control systems," IRE Trans. on Automatic Control, vol. AC-3, pp. 102–108; December, 1958.
[3] P. Eykhoff and O. J. M. Smith, "Optimalizing Control with Process-Dynamics Identification," to be published.
[4] G. E. P. Box, "Some General Considerations in Process Optimization," Dept. Math., Princeton University, Princeton, N. J., Tech. Rept. 13; April, 1958.
[5] B. Widrow, "Adaptive sampled-data systems—a statistical theory of adaption," 1959 IRE WESCON Convention Record, pt. 4, pp. 74–85.
[6] P. R. Stormer, "Adaptive or self-optimizing control systems—a bibliography," IRE Trans. on Automatic Control, vol. AC-4, pp. 65–69; May, 1959.
[7] W. C. Schultz and V. C. Rideout, "The selection and use of servo performance criteria," *Trans. AIEE (Applications and Industry,)* vol. 34, pp. 383–387; January, 1958.
[8] J. S. Bendat, "Principles and Applications of Random Noise Theory," John Wiley and Sons, Inc., New York, N. Y.; 1958.
[9] W. B. Davenport, Jr., R. A. Johnson, and D. Middleton, "Statistical errors in measurements on random time functions," *J. Appl. Phys.*, vol. 23, pp. 377–388; April, 1952.
[10] R. J. McGrath, "An Analytic and Simulator Study of an Adaptive Control System," Ph.D. dissertation, University of Wisconsin, Madison; 1960.
[11] G. C. Newton, L. A. Gould, and J. F. Kaiser, "Analytical Design of Linear Feedback Controls," John Wiley and Sons, Inc., New York, N. Y.; 1957.
[12] W. C. Schultz and V. C. Rideout, "Performance Measures: Past, Present, and Future," paper presented at the 7th Region IRE Conf., Seattle, Wash.; May, 1960.
[13] J. Lefkowity and D. P. Echman, "A review of optimizing/computer control," *Proc. Self Adaptive Flight Control Sys. Symp.*, WADC Tech. Rept. 59-49, pp. 181–197; March, 1959.
[14] M. Margolis and C. T. Leondes, "A parameter tracking servo for adaptive control systems," 1959 IRE WESCON Convention Record, pt. 4, pp. 104–115.
[15] G. W. Smith and B. E. Henderson, "Self-Adaptive Autopilot for Elastic Missiles," Martin Co., Denver, Colo., Res. Rept. R-60-11; June, 1960.
[16] T. Kitamori, "Applications of orthogonal functions to the determination of process dynamic characteristics and to the construction of self-optimizing control systems," *IFAC Congress*, vol. 1, pp. 82–86; July, 1960.
[17] A. A. Feldbaum, "Problems in the statistical theory of systems of automatic optimization," *IFAC Congress*, vol. 4, pp. 2088–2094; July, 1960.
[18] J. C. West, "Gain-modulated control systems," *IFAC Congress* vol. 3, pp. 1283–1287; July, 1960.

Optimum Prediction with a Mean Weighted Square Error Criterion*

CLARENCE C. GLOVER†, MEMBER, IRE

Summary—**The linear prediction theory is examined using a mean weighted square error criterion. A specific nondeterministic weighting function is used. The problem is reduced to that of solving integral equations which are written in terms of correlation functions which can be calculated by averaging over the ensemble. A complete solution is given for the problem using Gaussian statistics with no correlation between noise and true signal.**

Introduction

THE THEORY of linear least square smoothing and prediction was first developed by Wiener[1] and Kolmogorof.[2] Later it was examined in the frequency domain by Bode and Shannon.[3] Zadeh and Ragazzini[4] extended the theory to the case of finite observation time. Booton[5] and Davis[6] further extended the theory to time-varying systems with nonstationary statistical inputs. In all of this work the system was considered optimum when the least mean-square error was obtained.

The criterion of least mean square error has the disadvantage that large errors are weighted quite heavily even when they occur at a time when the variable under consideration is large. This is usually contrary to what is desired. For example, an error of one foot in a measurement of 100 miles is not nearly as disturbing as is the same error in a measurement of 100 feet. Usually the per cent of value of error is of more interest than is the absolute error. This suggests using the mean square value of the quotient obtained by dividing the error by the value of the variable as the criterion of performance.

This approach proves untenable when the variable has both positive and negative values and assumes the value of zero for some instant of time. Any error which occurred when the variable assumed the value of zero would be weighted infinitely large. Therefore, though one is in general interested in per cent error, in all measurements there is some value which should be considered as essentially zero. This dictates the lowest absolute accuracy of interest and suggests the following criterion of performance:

$$k = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \frac{[f(t) - f_d(t)]^2}{[f_d(t)]^2 + \delta^2}\, dt \tag{1}$$

where $f(t)$ is the obtained function of time and $fd(t)$ is the desired function of time. δ^2 dictates the lowest absolute accuracy of interest. The purpose of this paper is to examine how the linear prediction theory would be altered by adopting the latter criterion. This criterion shall be referred to as the mean weighted square error.

The idea of a mean weighted square error criterion has previously been introduced in the literature by Murphy and Bold,[7] but in their paper it was weighted by a deterministic function of t alone. In this paper, the weighting is by a nondeterministic function, *i.e.*,

$$\left[\frac{1}{[f_d(t)]^2 + \delta^2}\right].$$

Some of the mathematics involved in the two papers are very similar, but the two processes are actually physically quite different. All the correlation functions in this paper are such that they are easily calculated from ensemble averages. The entire paper is limited to cases where the statistics are time stationary. The stochastic process involved is ergodic. All functions and variables which are in the time domain are real functions of real variables. Due to the similarity of the problems treated, part of the notation used in this paper has been adopted from Murphy and Bold.

Notation

In order to avoid repeated writing of certain integrals, the following notation will be introduced:

$$\overline{f(t)} \equiv \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f(t)dt. \tag{2}$$

In general, $f(t)$ will be restricted so that

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f(t)dt = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f(t+\tau)dt. \tag{3}$$

Using this notation, the mean-square error can be written as $\overline{[e(t)]^2}$ and the mean-weighted-square error

* Received by the PGAC, June 15, 1960; revised manuscript received November 10, 1960.

† Dept. of Elec. Engrg., The Johns Hopkins University, Baltimore, Md.

[1] N. Wiener, "The Interpolation, Extrapolation, and Smoothing of Stationary Time Series," John Wiley and Sons, Inc., New York, N. Y.; 1949.

[2] A. Kolmogoroff, "Interpolation und extrapolation von stationären zufälligen folgen," *Bull. Acad. Sci. (URSS)*, Ser. Math. 4, pp. 3–14; 1941.

[3] H. W. Bode and C. E. Shannon, "A simplified derivation of linear least square smoothing and prediction theory," PROC. IRE, vol. 38, pp. 417–425; April, 1950.

[4] L. A. Zadeh and J. R. Ragazzini, "An extension of Wiener's theory of prediction," *J. Appl. Phys.*, vol. 21, pp. 645–655; July, 1950.

[5] R. C. Booton, Jr., "An optimization theory for time-varying linear systems with nonstationary statistical inputs," PROC. IRE, vol. 40, pp. 977–981; August, 1952.

[6] R. C. Davis, "On the theory of prediction of nonstationary stochastic processes," *J. Appl. Phys.*, vol. 23, pp. 1047–1053; September, 1952.

[7] G. J. Murphy and N. T. Bold, "Optimization based on a square error criterion with an arbitrary weighting function," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 24–30; January, 1960.

can be written as $\overline{[e(t)]^2W(t)}$. In the latter case $W(t)$ will be understood to be the nondeterministic function

$$\frac{1}{[f_d(t)]^2+\delta^2}.$$

In addition, the ordinary cross correlation function can be written as

$$\phi_{AB}(\tau_1,\tau_2)=\overline{A(t+\tau_1)B(t+\tau_2)}.$$

As this particular correlation function can be shown to be a function of a single variable $\tau=\tau_1-\tau_2$ it can also be written as $\phi_{AB}(\tau)$. Weighted correlation functions will be written as follows:

$$\phi_{WAB}(\tau_1,\tau_2)=\overline{A(t+\tau_1)B(t+\tau_2)W(t)}, \tag{4}$$

the functions $A(t)$ and $B(t)$ as well as $W(t)$ being bounded nondeterministic functions. In general, $\phi_{WAB}(\tau_1,\tau_2)$ cannot be expressed as a function of a single variable. The weighted auto correlation function commutates in the following sense:

$$\phi_{WAA}(\tau_1,\tau_2)=\phi_{WAA}(\tau_2,\tau_1).$$

The Problem

The problem is that of determining the optimum linear transfer function for a predicting filter where the filter will be considered optimum when the mean weighted square error is made a minimum.

The Convolution Theorem

As the systems under consideration are restricted to linear systems, the output $c(t)$ can be calculated from $g(t)$, the impulse response of the system, and $s(t)$, the input signal to the system, by the convolution theorem. Therefore:

$$c(t)=\int_{-\infty}^{\infty}g(x)s(t-x)dx.$$

Restricting the system to a physically realizable system dictates that $g(t)=0$ for $t<0$ and thus, the integral is

$$c(t)=\int_{0}^{\infty}g(x)s(t-x)dx. \tag{5}$$

The Weighted Square Error and Mean Weighted Square Error

The input signal $s(t)$ will be considered to have a true signal component $r(t)$ and a perturbing noise component $n(t)$. Thus, $s(t)=r(t)+n(t)$. For a predicting filter the desired output $c_d(t)=r(t+\tau)$, where τ is the prediction time. Then the error

$$\begin{aligned}e(t)&=c(t)-c_d(t)=c(t)-r(t+\tau)\\&=\int_0^{\infty}g(x)s(t-x)dx-r(t+\tau).\end{aligned}$$

The weighting function $W(t)$ would then be

$$\frac{1}{[r(t+\tau)]^2+\delta^2}$$

and, therefore, the following expression can be written for the weighted square error:

$$\begin{aligned}[e(t)]^2W(t)=&\int_0^{\infty}\int_0^{\infty}g(x)g(y)\left[\frac{s(t-x)s(t-y)}{[r(t+\tau)]^2+\delta^2}\right]dxdy\\&-2\int_0^{\infty}g(x)\left[\frac{s(t-x)r(t+\tau)}{[r(t+\tau)]^2+\delta^2}\right]dx\\&+\left[\frac{r(t+\tau)r(t+\tau)}{[r(t+\tau)]^2+\delta^2}\right].\end{aligned} \tag{6}$$

In taking the time average to obtain the mean-weighted-square error it will be assumed that the order of integration can be interchanged which gives the following expression for $\overline{[e(t)]^2W(t)}$:

$$\begin{aligned}&\overline{[e(t)]^2W(t)}\\&=\int_0^{\infty}\int_0^{\infty}g(x)g(y)\phi_{Wss}[-(x+\tau),-(y+\tau)]dxdy\\&\quad-2\int_0^{\infty}g(x)\phi_{Wsr}[-(x+\tau),0]dx+\phi_{Wrr}(0,0).\end{aligned} \tag{7}$$

Minimization of the Mean Weighted Square Error

The usual calculus of variation methods establish that for $\overline{[e(t)]^2W(t)}$ to be a minimum it is necessary that $g(x)$ satisfy the integral

$$\begin{aligned}&\int_0^{\infty}g(y)\phi_{Wss}[-(x+\tau),-(y+\tau)]dy\\&\qquad=\phi_{Wsr}[-(x+\tau),0]\;x\geq 0.\end{aligned} \tag{8}$$

The verisimilitude of this statement can be made apparent by the following procedure:

It is assumed that $g(x)$ is the function which gives the minimum value of $\overline{[e(t)]^2W(t)}$ equal to k. Then this impulse function is perturbed by adding an arbitrary function $h(x)$ multiplied by a constant to give $g'(x)=g(x)+\sigma h(x)$. This gives a new value to $\overline{[e(t)]^2W(t)}$ equal to k' which cannot be smaller than k. Thus, $k'-k\geq 0$. Deriving the expression for $k'-k$ in terms of $g(x)$, σ, and $h(t)$ gives the relation

$$\begin{aligned}k'-k=&2\sigma\int_0^{\infty}h(x)\left\{\int_0^{\infty}g(y)\phi_{Wss}[-(x+\tau),\right.\\&\left.-(y+\tau)]dy-\phi_{Wsr}[-(x+\tau),0]\right\}dx\\&+\sigma^2\int_0^{\infty}\int_0^{\infty}h(x)h(y)\phi_{Wss}[-(x+\tau),-(y+\tau)]dxdy.\end{aligned} \tag{9}$$

In order for this expression to ≥ 0 for all σ and $h(x)$,

(8) must be satisfied and, in addition, (10) must be satisfied.

$$\int_0^\infty \int_0^\infty h(x)h(y)\phi_{Wss}[-(x+\tau), -(y+\tau)]dxdy \geq 0 \quad (10)$$

As $W(t)$ is always >0, (10) is always satisfied when the order of integration can be interchanged as assumed in deriving (7).[8]

Extensions to the Case of Finite Memory

Considering a further restriction on $g(x)$ which demands that $g(x)=0$, $x>T$, extends the theory to that of finite observation time. Then the following condition must be fulfilled by $g(x)$: $g(x)=0$ for $x<0$ and $x>T$ and for $0\leq x\leq T$ (11) must be satisfied.

$$\int_0^T g(y)\phi_{Wss}[-(x+\tau), -(y+\tau)]dy = \phi_{Wsr}[-(x+\tau), 0]. \quad (11)$$

Eq. (12) is also a necessary condition.

$$\int_0^T \int_0^T h(x)h(y)\phi_{Wss}[-(x+\tau), -(y+\tau)]dxdy \geq 0. \quad (12)$$

Eq. (11) can be written in the following form:

$$\int_a^b g(y)k(x, y)dy = f(x) \quad (13)$$

where

$$k(x, y) = \phi_{Wss}[-(x+\tau), -(y+\tau)],$$

$f(x)=\phi_{Wsr}[-(x+\tau), 0]$, $g(y)$ is the unknown function, $a=0$, and $b=T$. This integral equation is the well-known linear integral equation of the first kind, where the kernel function $k(x,y)$ has the property that $k(x, y)=k(y, x)$ and,

$$\int_a^b \int_a^b g(x)g(y)k(x, y)dxdy \geq 0.$$

The solution of this mathematical problem is well treated in the mathematical literature. In addition this identical mathematical problem arises when the ordinary mean square error criterion is used with a nonstationary stochastic process. Davis[6] treats the problem in considerable detail. Therefore, the solution of our problem is essentially complete for the case of finite observation time.

The Integral with Infinite Limits

The methods of solution presented by Davis break down when the upper limit of integration is extended to ∞ as in (8). Therefore, we cannot consider the problem solved when the infinite past is used. This problem also arises in Booton's paper[5] which treats the problem of the optimization theory for time-varying linear systems with nonstationary statistical input.

[8] *Ibid.*, see steps 56–62.

Wiener-Hopf Integral Equation

When the kernel function can be expressed in a function of a single variable $(x-y)$, the integral equation is easily solved even when the upper limit is extended. In fact this form of the equation is the well-known Wiener-Hopf integral equation which can be written in the form

$$\int_0^\infty g(y)\phi_{ss}(x-y)dy = \phi_{sr}(x+\tau) \qquad x \geq 0. \quad (14)$$

This equation occurs when $W(t)$ is allowed to be equal to 1. Therefore, solving for $g(y)$ gives the optimum linear impulse response for the predicting filter when the mean-square-error criterion is used. Wiener's solution to this problem is well known and can be written as

$$G(j\omega) = \frac{1}{\Phi_{ss}^{+}(j\omega)} F\begin{bmatrix} t\geq 0 & F^{-1}\left\{e^{j\omega\tau}\dfrac{\Phi_{sr}(j\omega)}{\Phi_{ss}^{-}(j\omega)}\right\} \\ t<0 & 0 \end{bmatrix}$$

where $\Phi_{ss}^{+}(j\omega)\times\Phi_{ss}^{-}(j\omega)=\Phi_{ss}(j\omega)$, $G(j\omega)$ is the Fourier transform of $g(t)$, *i.e.* $G(j\omega)=F\{g(t)\}$. Similarly $\Phi_{ss}(j\omega)=F\{\phi_{ss}(t)\}$, etc. $\Phi_{ss}(j\omega)$ is factored so that $\Phi_{ss}^{+}(s)$ has no poles or zeros with their real part greater than zero, and $\Phi_{ss}^{-}(s)$ has no poles or zeros with their real part less than zero. Poles and zeros with their real parts zero are split between $\Phi_{ss}^{+}(j\omega)$ and $\Phi_{ss}^{-}(j\omega)$.

This gives a special niceness to using the mean square error criterion, as it leads to an integral equation which is readily solved by taking Fourier transforms (the solution being in terms of the ordinary correlation functions or their Fourier transforms).

Solution for Specified Statistics

The use of the mean weighted square error criterion would be much more tenable if a solution for the optimum network could also be found in terms of the ordinary correlation function instead of the special correlation functions that appear in (8). Such a solution would depend on the nature of the process generating $r(t)$, *i.e.*, the kind of statistical distribution involved. If the nature of the statistical process is known, then ϕ_{WAB} can be found by the following relation between time and ensemble averages:

$$\phi_{WAB}(\tau_1, \tau_2) = \iiint_{-\infty}^{+\infty} \frac{yz}{x^2+\delta^2} p(x, y, z)dydzdx \quad (15)$$

where $p(x, y, z)$ is the joint probability density function of the random variables $x=f_d(t)$, $y=A(t+\tau_1)$, and $z=B(t+\tau_2)$. It is necessary to specify the nature of the process in order to proceed to a more detailed solution. One type process of special importance is the Gaussian process.

The Solution for Gaussian Process

Considering the process generating $r(t)$ as a Gaussian process, (15) can now be written in terms of the multi-

variate Gaussian distribution[9] for both of the weighted correlation functions needed in (8). Then the indicated integration can be performed, which will allow the two weighted correlation functions to be expressed in terms of ordinary correlation functions. It is convenient to use the normalized correlation function ρ.

$$\rho_{AB} = \frac{\phi_{AB} - m_A m_B}{\sigma_A \sigma_B} \quad \text{where} \quad m_A = \overline{A(t)},$$

$$\sigma_A = \sqrt{\overline{[A(t) - m_A]^2}}, \text{ etc.}$$

An Example Problem

To illustrate the procedure the following simplified problem is considered, one of pure prediction [*i.e.*, $n(t) = 0$], where the coordinate system has been selected so that $\overline{r(t)} = 0$. Then the integral equation which must be solved will have the form

$$\int_0^\infty g(y)\phi_{Wrr}[-(x+\tau), -(y+\tau)]dy = \phi_{Wrr}[-(x+\tau), 0] \qquad x \geq 0. \quad (16)$$

Thus, it is necessary to find $\phi_{Wrr}(\tau_1, \tau_2)$ in terms of ρ_{rr}. Substituting the proper probability density function into (15) and performing the integration gives the following results:

$$\phi_{Wrr}(\tau_1, \tau_2) = \rho_{rr}(\tau_1)\rho_{rr}(\tau_2) \cdot \left\{1 - \sqrt{\frac{\pi}{2}}\left(\frac{\delta}{\sigma_r} + \frac{\sigma_r}{\delta}\right)\exp\frac{1}{2}\left(\frac{\delta}{\sigma_r}\right)^2 \operatorname{erfc}\frac{1}{\sqrt{2}}\left(\frac{\delta}{\sigma_r}\right)\right\} + \rho_{rr}(\tau_1 - \tau_2)\sqrt{\frac{\pi}{2}}\left(\frac{\sigma_r}{\delta}\right)\exp\frac{1}{2}\left(\frac{\delta}{\sigma_r}\right)^2 \operatorname{erfc}\frac{1}{\sqrt{2}}\left(\frac{\delta}{\sigma_r}\right) \quad (17)$$

where erfc is the complementary function to the error function (erfc $x = 1 - \operatorname{erf} x$).[10]

The normalized auto correlation function $\rho_{AA}(\tau)$ has the following properties: $\rho_{AA}(0) = 1$, $\rho_{AA}(\tau) = \rho_{AA}(-\tau)$. Thus, (16) can be written as

$$\int_0^\infty g(y)\left\{\rho_{rr}(x+\tau)\rho_{rr}(y+\tau) \cdot \left[1 - \sqrt{\frac{\pi}{2}}\left(\frac{\delta}{\sigma_r} + \frac{\sigma_r}{\delta}\right)\exp\frac{1}{2}\left(\frac{\delta}{\sigma_r}\right)^2 \operatorname{erfc}\frac{1}{\sqrt{2}}\left(\frac{\delta}{\sigma_r}\right)\right] + \rho_{rr}(x-y)\sqrt{\frac{\pi}{2}}\left(\frac{\sigma_r}{\delta}\right)\exp\frac{1}{2}\left(\frac{\delta}{\sigma_r}\right)^2 \operatorname{erfc}\frac{1}{\sqrt{2}}\left(\frac{\delta}{\sigma_r}\right)\right\} dy = \rho_{rr}(x+\tau)\left[1 - \sqrt{\frac{\pi}{2}}\left(\frac{\delta}{\sigma_r}\right)\exp\frac{1}{2}\left(\frac{\delta}{\sigma_r}\right)^2 \operatorname{erfc}\frac{1}{\sqrt{2}}\left(\frac{\delta}{\sigma_r}\right)\right] \qquad x \geq 0. \quad (18)$$

[9] Davenport and Root, "An Introduction to the Theory of Random Signals and Noise," McGraw-Hill Book Co., Inc., New York, N. Y., pt. 8-3, pp. 151–153; 1958.

[10] "Mathematical tables," in "Handbook of Chemistry and Physics," Chemical Rubber Publ. Co., Cleveland, Ohio, 11th ed., p. 301; 1959.

This can be simplified by letting

$$\sqrt{\frac{\pi}{2}}\exp\frac{1}{2}\left(\frac{\delta}{\sigma_r}\right)^2 \operatorname{erfc}\frac{1}{\sqrt{2}}\left(\frac{\delta}{\sigma_r}\right) = A,$$

and then

$$\frac{A}{\frac{\delta}{\sigma_r}\left(1 - \frac{\delta}{\sigma_r}A\right)} = B.$$

Then (16) can be written as

$$\rho_{rr}(x+\tau) = \frac{B}{1 - (1-B)\int_0^\infty g(y)\rho_{rr}(y+\tau)dy \cdot \int_0^\infty g(y)\rho_{rr}(x-y)dy} \qquad x \geq 0. \quad (19)$$

For $\tau \leq 0$ the solution to (19) is simply $G(j\omega) = e^{j\omega\tau}$ or $g(t) = \delta(t-\tau)$ which means the noiseless signal need only be delayed by $|\tau|$ seconds. For $\tau > 0$ several solutions were obtained using particular forms of $\rho_{rr}(\tau)$. In each problem examined the solution proved to differ from that obtained using the mean square error criterion by a constant multiplying factor.

This leads directly to the following general solution for the problem of pure prediction using the mean weighted square error criterion where the signal $r(t)$ is generated by a Gaussian process. The problem is first solved using the ordinary mean square error criterion which can be readily accomplished by using Wiener's solution. This yields a solution $g_1(t)$ for the impulse function of the system which is a solution of the following Wiener-Hopf equation:

$$\int_0^\infty g_1(y)\rho_{rr}(x-y)dy = \rho_{rr}(x+\tau) \qquad x \geq 0. \quad (20)$$

Then the solution $g(t)$ to the problem using the mean weighted square error can be found by multiplying $g_1(t)$ by a constant $1/K$ (*i.e.*, $g(t) = (1/K)g_1(t)$), where K is given by

$$K = B + (1-B)\int_0^\infty g_1(y)\rho_{rr}(y+\tau)dy. \quad (21)$$

To prove that this solution holds, it must be shown that (19) is satisfied by $g(t)$, when $g_1(y)$ satisfies (20) and $Kg(y) = g_1(y)$. These conditions allow (20) to be rewritten as follows:

$$\int_0^\infty g(y)\rho_{rr}(x-y)dy = \frac{1}{K}\rho_{rr}(x+\tau) \qquad x \geq 0. \quad (22)$$

Substituting the right side of (22) for the left side in

(19) and substituting $(1/K)g_1(t)$ for $g(t)$ gives

$$\rho_{rr}(x+\tau) = \frac{B(1/K)\rho_{rr}(x+\tau)}{1-(1-B)(1/K)\int_0^\infty g_1(y)\rho_{rr}(y+\tau)dy} \qquad x \geq 0. \quad (23)$$

Therefore,

$$1 = \frac{B}{K-(1-B)\int_0^\infty g_1(y)\rho_{rr}(y+\tau)dy}$$

and

$$K = B + (1-B)\int_0^\infty g_1(y)\rho_{rr}(y+\tau)dy. \qquad Q.E.D.$$

Extension to Other Problems

This solution can easily be extended to cases where $n(t) \neq 0$ providing that $\overline{n(t)} = 0$ and that there is no cross correlation between the noise $n(t)$ and the true signal $r(t)$. The equation corresponding to (19) is as follows:

$$\rho_{rr}(x+\tau) = \frac{B}{1-(1-B)\int_0^\infty g(y)\rho_{rr}(y+\tau)dy} \cdot \int_0^\infty g(y)\rho_{ss}(x-y)dy, \qquad x \geq 0 \quad (24)$$

where

$$\rho_{ss} = \left(\frac{\sigma_r}{\sigma_s}\right)^2 \rho_{rr} + \left(\frac{\sigma_n}{\sigma_s}\right)^2 \rho_{nn}.$$

Thus the solution $g(t) = (1/K)g_1(t)$ is found by solving the following Wiener-Hopf integral equation to obtain $g_1(t)$:

$$\int_0^\infty g_1(y)\rho_{ss}(x-y)dy = \rho_{rr}(x+\tau), \qquad x \geq 0 \quad (25)$$

and

$$K = B + (1-B)\int_0^\infty g_1(y)\rho_{rr}(y+\tau)dy.$$

A and B are exactly as in the previous problem.

Attempts to obtain similar solutions when the noise and true signal are correlated and with other types of statistics lead to the necessity of solving higher order transcendental equations and probably can only be solved with recourse to machine computations.

Conclusion

The statistical optimization theory developed by Wiener has been extended to using a mean weighted square error criterion. In the case of one specific weighting function, the general problem has been reduced to one of solving integral equations. A complete solution is given for the problem using Gaussian statistics with no correlation between noise and true signal. The techniques employed can be extended to other weighting functions providing the function $W(t)$ is always greater than 0 and is a nondeterministic function of time as indicated in the discussion.

The latter restriction is important because when the weighting function is a definite function of time, certain integrations involved may not exist in the sense that they yield definite functions but yield new random variables as discussed by Davenport and Root.[11] For example $\phi_{WAB}(\tau_1, \tau_2)$ may be a random variable. In other cases it may reduce to $\phi_{AB}(\tau_2 - \tau_1)$ to give exactly the same solution as $W(t) = 1$ gives.

Appendix

Illustrative Problem

For an example of a particular problem we consider problem number one.[12] Let $s(t)$ be a signal which is a sample function from a stationary random process with spectral density

$$\Phi(j\omega) = \frac{1}{1+(j\omega)^4}.$$

Show that the least mean square error predicting filter which operates on the infinite past has system function

$$G_1(j\omega) = e^{-\tau/\sqrt{2}}\left[\cos\frac{\tau}{\sqrt{2}} + \sin\frac{\tau}{\sqrt{2}} + \sqrt{2}\,(j\omega)\sin\frac{\tau}{\sqrt{2}}\right]$$

where τ is the lead time. First, we shall solve the problem as stated and then we shall solve the problem for the least mean weighted square error, assuming that the statistics are Gaussian.

Solution to the Problem

$$\Phi(S) = \frac{1}{1+S^4} = \left(\frac{1}{1+\sqrt{2}S+S^2}\right)\left(\frac{1}{1-\sqrt{2}S+S^2}\right).$$

The poles of the left term are

$$S_* = \frac{-1+j}{\sqrt{2}} \quad \text{and} \quad S_* = \frac{-1-j}{\sqrt{2}}.$$

The poles of the right term are

$$S_* = \frac{1+j}{\sqrt{2}} \quad \text{and} \quad S_* = \frac{1-j}{\sqrt{2}}.$$

Thus

$$\Phi^+(S) = \left(\frac{1}{1+\sqrt{2}S+S^2}\right)$$

[11] Davenport and Root, *op. cit.*, see pp. 65–66.
[12] Davenport and Root, *op. cit.*, see p. 247.

and

$$\Phi^-(S) = \left(\frac{1}{1 - \sqrt{2}\,S + S^2}\right).$$

Then

$$\Phi^+(j\omega) = \left(\frac{1}{1 + \sqrt{2}\,(j\omega) + (j\omega)^2}\right)$$

and

$$\Phi^-(j\omega) = \left(\frac{1}{1 - \sqrt{2}\,(j\omega) + (j\omega)^2}\right).$$

Note that $\Phi^+(j\omega)$ and $\Phi^-(j\omega)$ are complex conjugates in the sense that $\Phi^+(-j\omega) = \Phi^-(j\omega)$. Following the procedure indicated after (14) we need to find the inverse Fourier transform of $\Phi^+(j\omega)e^{j\omega\tau}$ for $t > 0$. For this particular problem we can accomplish this by taking the inverse Laplace transform of $\Phi^+(S)$ and then replacing the variable t with a new variable $t+\tau$ as follows:

$$t > 0 \qquad F^{-1}[\Phi^+(j\omega)e^{j\omega\tau}] = L^{-1}[\Phi^+(S)]\,\big|_{t\to t+\tau}$$

$$= L^{-1}\left[\frac{1}{1 + \sqrt{2}\,S + S^2}\right]\Bigg|_{t\to t+\tau} = \sqrt{2}\,e^{-t/\sqrt{2}}\sin\frac{t}{\sqrt{2}}\Bigg|_{t\to t+\tau}$$

$$= \sqrt{2}\,e^{-\tau/\sqrt{2}}\left[\sin\frac{\tau}{\sqrt{2}}\left(e^{-t/\sqrt{2}}\cos\frac{t}{\sqrt{2}}\right) + \cos\frac{\tau}{\sqrt{2}}\left(e^{-t/\sqrt{2}}\sin\frac{t}{\sqrt{2}}\right)\right].$$

To complete the process we take the Laplace transform of the above expression and multiply by $1/\Phi^+(S)$ to obtain

$$G_1(S) = e^{-\tau/\sqrt{2}}\left[\sin\frac{\tau}{\sqrt{2}} + \cos\frac{\tau}{\sqrt{2}} + \left(\sqrt{2}\sin\frac{\tau}{\sqrt{2}}\right)(S)\right]$$

or

$$G_1(j\omega) = e^{-\tau/\sqrt{2}}\left[\sin\frac{\tau}{\sqrt{2}} + \cos\frac{\tau}{\sqrt{2}} + \left(\sqrt{2}\sin\frac{\tau}{\sqrt{2}}\right)(j\omega)\right]$$

To obtain the solution for the least mean weighted square error we need to find $\phi(t) = F^{-1}[\Phi(j\omega)]$. For this particular problem we can accomplish this by breaking $\Phi(S)$ into the sum of two terms, where the first term has poles and zeros in the left-hand plane and the second term has poles and zeros in the right-hand plane. Then we take the inverse Laplace transform of the first and replace the variable t with a new variable $|t|$ as follows:

$$\Phi(S) = \frac{1}{1 + S^4} = \left(\frac{\frac{1}{2} + \frac{1}{2\sqrt{2}}S}{1 + \sqrt{2}\,S + S^2}\right) + \left(\frac{\frac{1}{2} - \frac{1}{2\sqrt{2}}S}{1 - \sqrt{2}\,S + S^2}\right),$$

$$L^{-1}\left[\frac{\frac{1}{2} + \frac{1}{2\sqrt{2}}S}{1 + \sqrt{2}\,S + S^2}\right]\Bigg|_{t\to|t|} = \frac{1}{2\sqrt{2}}\,e^{-|t|/\sqrt{2}}\left(\cos\frac{|t|}{\sqrt{2}} + \sin\frac{|t|}{\sqrt{2}}\right) = \phi(t).$$

Thus,

$$\rho(t) = e^{-|t|/\sqrt{2}}\left(\cos\frac{|t|}{\sqrt{2}} + \sin\frac{|t|}{\sqrt{2}}\right)$$

and $\sigma^2 = 1/2\sqrt{2}$. We also need $g_1(t) = F^{-1}[G_1(j\omega)]$ which gives

$$g_1(t) = e^{-\tau/\sqrt{2}}\left[\left(\sin\frac{\tau}{\sqrt{2}} + \cos\frac{\tau}{\sqrt{2}}\right)\delta(t) + \left(\sqrt{2}\sin\frac{\tau}{\sqrt{2}}\right)\delta'(t)\right]$$

where $\delta(t)$ is the unit impulse function and $\delta'(t)$ is the first derivative of the unit impulse function. These functions are being used in the manner described in Appendix 1,[13] (*i.e.*, $F[\delta^n(t)] = (j\omega)^n$ and

$$\int_{x_0-\epsilon}^{x_0+\epsilon} f(x)\delta^{(n)}(x - x_0)dx = (-1)^n\frac{d^n}{dx^n}f(x)\Big|_{x=x_0}.$$

Thus, we can write an expression for the constant K using (21).

$$K = B + (1 - B)\int_0^\infty e^{-\tau/\sqrt{2}} \cdot \left[\left(\sin\frac{\tau}{\sqrt{2}} + \cos\frac{\tau}{\sqrt{2}}\right)\delta(t) + \left(\sqrt{2}\sin\frac{\tau}{\sqrt{2}}\right)\delta'(t)\right] \cdot \left[e^{-|t+\tau|/\sqrt{2}}\left(\cos\frac{|t+\tau|}{\sqrt{2}} + \sin\frac{|t+\tau|}{\sqrt{2}}\right)\right]dt.$$

Evaluating gives

$$K = B + (1 - B)\left\{e^{-2\tau/\sqrt{2}}\left[\left(\cos\frac{\tau}{\sqrt{2}} + \sin\frac{\tau}{\sqrt{2}}\right)^2 + 2\sin^2\frac{\tau}{\sqrt{2}}\right]\right\},$$

and then we can write an expression for

$$g(t) = \frac{1}{K}\,g_1(t) \quad \text{or} \quad G(j\omega) = \frac{1}{K}\,G_1(j\omega).$$

[13] Davenport and Root, *op. cit.*, see pp. 365–370.

Control Systems with Minimum Spectral Bandwidth of Plant Input*

JAMES C. HUNG†, MEMBER, IRE

Summary—A design method for control systems that minimizes the spectral bandwidth of the plant input signal is discussed. The plant input signal is minimized subject to the constraint that the integral square error for deterministic inputs or the mean square error for random inputs be limited to a known desired value.

The control system transfer function that satisfies these requirements is derived, and the functions used in bandwidth shaping are discussed. An example of a system design using this technique is given.

General

THE design of servo systems to minimize the high-frequency power of the plant-input is derived in this paper. The minimization is accomplished with a constraint of a given integral square error for deterministic inputs or mean square error for random inputs.

Limitation of the high-frequency power of the plant-input avoids exciting the higher-frequency modes of the plant. The transfer function of a plant can be separated into two parts, the low-frequency modes and the high-frequency modes. The low-frequency modes are simple in form and well-defined. The high-frequency modes are usually not known exactly, and inclusion of these terms in the design requires complicated compensation networks in the control loop. Examples of high-frequency plant modes are motor shaft modes, body resonant modes of missiles, etc.

The design of control systems to minimize the bandwidth of the closed-loop transfer function subject to an error constraint has appeared in the literature.[1] With certain control plants it is possible that limiting the closed-loop system bandwidth does not limit the spectral bandwidth of the plant input, and, as a result, the high-frequency modes of the plant are nevertheless excited. The conditions for which the system bandwidth limitation does not also limit the plant-input signal bandwidth can be simply developed.

Let $G(s)$ be the known part of the plant transfer function whose denominator polynomial is h degree higher than its numerator polynomial, and let $K(s)$ represent the closed-loop transfer function whose denominator is k degree higher than its numerator polynomial. $K(s)$ declines as frequency tends to infinity with an asymptotic slope of $-6k$ db per octave, and $G(s)$ declines with a slope of $-6h$ db per octave. With reference to Fig. 1, the transfer function relating the plant input to the system input is given by

$$M(s) = \frac{I(s)}{R(s)} = \frac{K(s)}{G(s)} . \tag{1}$$

Eq. (1) shows that at high frequency, the asymptotic response of $M(s)$ varies at a slope of $-6(k-h)$ db per octave, which is dependent on the values of k and h. If h is greater than or equal to k the function $M(s)$ does not decline at all outside the system bandwidth. Under this condition, the high-frequency input and noise is not restricted from entering the plant, and the high-frequency modes of the plant can still be excited.

To cope with this difficulty, this paper proposes a different approach to the problem. Instead of limiting the bandwidth of the closed-loop transfer function $K(s)$, the high-frequency power of the plant-input is minimized in the optimum design procedure. The system thus obtained is certain not to excite the resonant modes of the plant.

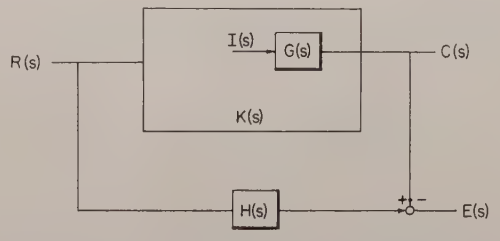


Fig. 1—Servo control system with fixed plant.

Theory

Assume that the transfer function $G(s)$ of the controlled plant is given, and that the tolerable integral square error, L, of the closed-loop system in response to a system deterministic input of $R(s)$ is specified. With the given quantities a closed-loop transfer function $K(s)$ is to be solved for, whose manipulated variable $I(s)$, the plant input, referring to Fig. 1, has minimum high-frequency power. The quantity to be minimized is

$$\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} I\bar{I}W\bar{W}\,ds = \min \tag{2}$$

subject to the constraint condition that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T [e(t)]^2 dt = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} E\bar{E}\,ds \leqq L. \tag{3}$$

* Received by the PGAC, May 10, 1960; revised manuscript received October 4, 1960. The research in this paper was supported by the USAF under contract No. AF 49(638)-586 monitored by the Office of Scientific Res., Air Res. and Dev. Command, Washington 25, D. C.

† Elec. Engrg. Dept., New York University, New York, N. Y.

[1] G. C. Newton, Jr., "Design of control system for minimum bandwidth," *Trans. AIEE*, vol. 74, pp. 161–168; July, 1955.

In (2), $W(s)$ is the weighting function which is used to shape the high-frequency asymptote of $I(s)$. $W(s)$ is in general a high pass filter of the form s^n with $n \geqq 0$. Since the output of the high pass filter is the quantity that is minimized, this forces the input signal to fall off at high frequencies. Thus, the high-frequency power of the plant-input is limited. The Laplace transform variables are represented by upper case letters, and their complex conjugates are represented by barred upper case letters. For example,

$$R = R(s), \qquad \overline{R} = R(-s).$$

Let $H(s)$ be the desired closed-loop transfer function. Then the error between the desired output and the system output is given by

$$E = R(H - K). \tag{4}$$

The plant-input signal is

$$I = \frac{K}{G} R. \tag{5}$$

Eqs. (2) and (3) may be combined into one integral that is to be minimized by using the technique of Lagrange multipliers.[2] With (4) and (5) substituted, the integral is

$$\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} R\overline{R}\left[\frac{K\overline{K}}{G\overline{G}} W\overline{W} + \lambda(H - K)(\overline{H} - \overline{K})\right] ds = \min \tag{6}$$

when λ is the Lagrange multiplier whose value is determined by L, the tolerable integral square error.

The integrand of (6) is

$$F = R\overline{R}\left[\frac{K\overline{K}W\overline{W}}{G\overline{G}} + \lambda(H - K)(\overline{H} - \overline{K})\right]. \tag{7}$$

The integral of (6) is to be minimized considering that the system transfer function, $K(s)$, is to be varied. It has been shown that the partial derivative of the integrand with respect to the conjugate of the varied function must have poles only in the right-half plane for the integral to be a minimum and for the system function to be physically realizable.[3] Let

$$X = \frac{\partial F}{\partial \overline{K}} = R\overline{R}\left[\frac{KW\overline{W}}{G\overline{G}} - \lambda(H - K)\right]$$

where X has only right-half plane poles. Rearranging,

$$X = R\overline{R}\left[K\left(\frac{W\overline{W}}{G\overline{G}} + \lambda\right) - \lambda H\right]. \tag{8}$$

Let

$$\frac{W\overline{W}}{G\overline{G}} + \lambda = Y\overline{Y} \tag{9}$$

and

$$R\overline{R} = Z\overline{Z} \tag{10}$$

where Y and Z have all their poles and zeros in the LHP, while $\overline{Y}$ and $\overline{Z}$ have all their poles and zeros in the RHP. Substituting (9) and (10) into (8) and dividing by $\overline{Y}\overline{Z}$ gives

$$YZK = \frac{\lambda ZH}{\overline{Y}} + \frac{X}{\overline{Y}\overline{Z}}. \tag{11}$$

The term in the left-hand side of (11) is analytic in the RHP, since K is implemented to be a stable system and Z and Y are analytic in the RHP by definition. The first term in the right-hand side of (11) may have poles in both sides of the s-plane. Using a partial fraction method, this term can be separated into two parts as

$$\frac{\lambda ZH}{\overline{Y}} = \left[\frac{\lambda ZH}{\overline{Y}}\right]_+ + \left[\frac{\lambda ZH}{\overline{Y}}\right]_-$$

where the part inside the bracket $[]_+$ is analytic in the RHP, and the part inside the bracket $[]_-$ is analytic in the LHP. The second term in the right-hand side of (11) is analytic in the LHP from the definitions of X, $\overline{Y}$ and $\overline{Z}$. Eq. (11) can therefore be written as

$$YZK - \left[\frac{\lambda ZH}{\overline{Y}}\right]_+ = \left[\frac{\lambda ZH}{\overline{Y}}\right]_- + \frac{X}{\overline{Y}\overline{Z}}. \tag{12}$$

The left-hand side of (12) is analytic in the RHP, while the right-hand side is analytic in the LHP. There are no poles common to both sides of (12); therefore, both sides should be equal to zero, giving

$$YZK - \left[\frac{\lambda ZH}{\overline{Y}}\right]_+ = 0. \tag{13}$$

The optimum closed-loop transfer function is

$$K = \frac{1}{YZ}\left[\frac{\lambda ZH}{\overline{Y}}\right]_+. \tag{14}$$

To find the value of λ, (14) and (4) are substituted into (3) and the integration is performed.

It is obvious that the use of different weighting functions, $W(s)$, results in plant-input transfer functions,

[2] R. Courant and D. Herbert, "Methods of Mathematical Physics," Interscience Publishing Co. Inc., New York, N. Y.; 1953.

[3] S. S. L. Chang, "Two network theorems for analytical determination of optimum response physically realizable network characteristics," Proc. IRE, vol. 43, pp. 1128–1135; September, 1955.

$M(s)$, having different high-frequency cutoff characteristics. This can be seen from the following consideration.

Consider a system with $H=1$, $R=1/s$, and $G=Z/P$ where Z is a polynomial of degree z, and P is a polynomial of degree p, and $p>z$. Then

$$G\overline{G} = \frac{Z\overline{Z}}{P\overline{P}}.$$

Using (9) and (10), we have

$$Y\overline{Y} = \frac{W\overline{W}}{G\overline{G}} + \lambda = \frac{W\overline{W}P\overline{P} + \lambda Z\overline{Z}}{Z\overline{Z}} = \frac{Q\overline{Q}}{Z\overline{Z}}$$

$$Z\overline{Z} = R\overline{R} = \frac{1}{S^2}$$

where W is equal to s^n, and Q is a polynomial of degree $q=p+n$. From (14),

$$K = \frac{1}{YZ}\left[\frac{\lambda Z}{\overline{Y}}\right]_+ = \frac{A}{Y} = A\frac{Z}{Q}$$

where A is a constant. Since $q-z=p+n-z$ is always positive, $K(s)$ always vanishes as frequency increases. The plant input function

$$M = \frac{K}{G} = A\frac{ZP}{QZ} = A\frac{P}{Q}$$

has a high-frequency cutoff rate of $q-p=p+n-p=n$ units. That is, M declines at the rate of $-6n$ db per octave at the high-frequency end. Techniques of choosing the value of n will be illustrated in the example following this section.

The method presented so far is for systems having deterministic inputs. Extension of the method to the case of random inputs is obvious. In the latter case, the input power spectrum Φ_{rr} replaces $R\overline{R}$ in (6), and the high-frequency power of the plant input is minimized subject to a specified mean square error constraint.

It should be noted that the same method can be used for the design of optimum closed-loop systems either by minimizing the integral square error, or by minimizing the mean square error, subject to a specified high-frequency plant input power limitation.

Illustration

To illustrate the design procedure discussed in the previous section, an example of the design of a closed-loop system will be given. Referring to Fig. 1, let the known part of the plant transfer function be $G=1/s^2$, the desired closed-loop function be $H(s)=1$, and the input $R(s)=1/s$. The tolerable integral square-error of the system in response to the deterministic input is assumed to be not greater than one, *i.e.*, $L\leqq 1$.

First let the weighting function be $W=1$. Then (9) gives

$$Y\overline{Y} = \frac{W\overline{W}}{G\overline{G}} + \lambda = s^4 + \lambda$$

$$= (s+a)(s+\bar{a})(s-a)(s-\bar{a}) \tag{15}$$

where

$$a = \lambda^{1/4}e^{j(\pi/4)} \quad \text{and} \quad \bar{a} = \lambda^{1/4}e^{-j(\pi/4)}. \tag{16}$$

Eq. (10) gives

$$ZZ = R\overline{R} = \frac{1}{S^2}. \tag{17}$$

Separating (15) into left- and right-half plane terms, we have

$$\left.\begin{aligned} Y &= (s+a)(s+\bar{a}) \\ \overline{Y} &= (s-a)(s-\bar{a}) \end{aligned}\right\}. \tag{18}$$

Substitute (18), R, and H into (14),

$$K = \frac{1}{YZ}\left[\frac{\lambda ZH}{\overline{Y}}\right]_+.$$

Evaluating using the left-half plane terms in the bracket

$$K = \frac{s}{(s+a)(s+\bar{a})}\frac{\lambda}{a\bar{a}s} = \frac{a\bar{a}}{(s+a)(s+\bar{a})}. \tag{19}$$

The plant input function is given by

$$M = \frac{K}{G} = \frac{s^2 a\bar{a}}{(s+a)(s+\bar{a})}. \tag{20}$$

To evaluate λ, substitute (19) into (3), giving

$$\int_0^\infty e^2 dt = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} E\overline{E}ds$$

$$= \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} R\overline{R}(1-K)(1-\overline{K})ds$$

$$= -\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} \frac{[s+(a+\bar{a})][s-(a+\bar{a})]}{(s+a)(s+\bar{a})(s-a)(s-\bar{a})}ds$$

$$= \left(\frac{1}{a+\bar{a}}\right)\left(\frac{a^2+3a\bar{a}+\bar{a}^2}{2a\bar{a}}\right).$$

From (16) it can readily be seen that

$$a^2 + \bar{a}^2 = 0.$$

So

$$\int_0^\infty e^2 dt = \frac{3}{2(a+\bar{a})} = \frac{1.06}{\lambda^{1/4}} \leq L.$$

Since $L=1$, the limiting value of λ is

$$\lambda = (1.06)^4 = 1.262. \tag{21}$$

Insert the value of λ into (19) and (20), resulting in

$$K = \frac{1.123}{(s+1.06/\underline{45})(s+1.06/\underline{-45})}$$

$$= \frac{1.123}{s^2+1.5s+1.123}, \tag{22}$$

and

$$M = \frac{1.123s^2}{s^2+1.5s+1.123}. \tag{23}$$

Next, let $W=s$, and follow the same procedure. For this case,

$$\lambda = 20.85, \tag{24}$$

$$K = \frac{4.566}{(s+1.659)(s+1.659/\underline{60})(s+1.659/\underline{-60})}$$

$$= \frac{4.566}{s^3+3.318s^2+5.5s+4.566}, \tag{25}$$

and

$$M = \frac{4.566s^2}{s^3+3.318s^2+5.5s+4.566}. \tag{26}$$

Finally, let $W=s^2$, obtaining

$$\lambda = 435.48, \tag{27}$$

$$K = \frac{26.46}{(s+2.268/\underline{67.5})(s+2.268/\underline{-67.5})(s+2.268/\underline{22.5})(s+2.268/\underline{-22.5})}$$

$$= \frac{26.46}{s^4+5.927s^3+17.56s^2+30.49s+26.46}, \tag{28}$$

and

$$M = \frac{26.46s^2}{s^4+5.927s+17.56s^2+30.49s+26.46}. \tag{29}$$

Fig. 2 shows the frequency response curve of the three plant-input functions given by (23), (26), and (29). The high-frequency asymptotes have slopes of 0 db, -6 db, and -12 db per octave for $W=1$, $W=s$, and $W=s^2$ respectively. From the curves it can be seen that if the plant high-frequency mode is lower than 4 radians per second, $W=1$ is the best weighting function to use. If the plant high-frequency mode is between 4 and 5 radians per second, then $W=s$ is the best weighting function. For modes higher than 5 radians per second, $W=s^2$ is the best one among the three. Using weighting functions with higher degrees than necessary certainly gives desirable high-frequency noise attenuation, but at the cost of an unnecessarily complex system. From the graph it can be seen that the higher, order weighting functions give systems that are more susceptible to noise disturbance near the cutoff region.

If series compensation alone were used for the above system, the compensation device would have a zero at the origin which would block off the dc reference input. This difficulty will be eliminated if both series and feedback compensation networks are used.

The system configuration to be used is shown in Fig. 3. The transfer function of the compensation units are given as follows. For the first system, $W=1$

$$G_s = 1.123$$

$$G_f = 1.33(s+0.75).$$

For the second system, $W=s$

$$G_s = \frac{4.566}{s+3.318}$$

$$G_f = 1.21(s+0.833).$$

For the third system, $W=s^2$

$$G_s = \frac{26.46}{s^2+5.927s+17.56}$$

$$G_f = 1.15(s+0.868).$$

The root-locus plots of these three systems are shown in Figs. 4, 5, and 6, and the corresponding gain points are indicated.

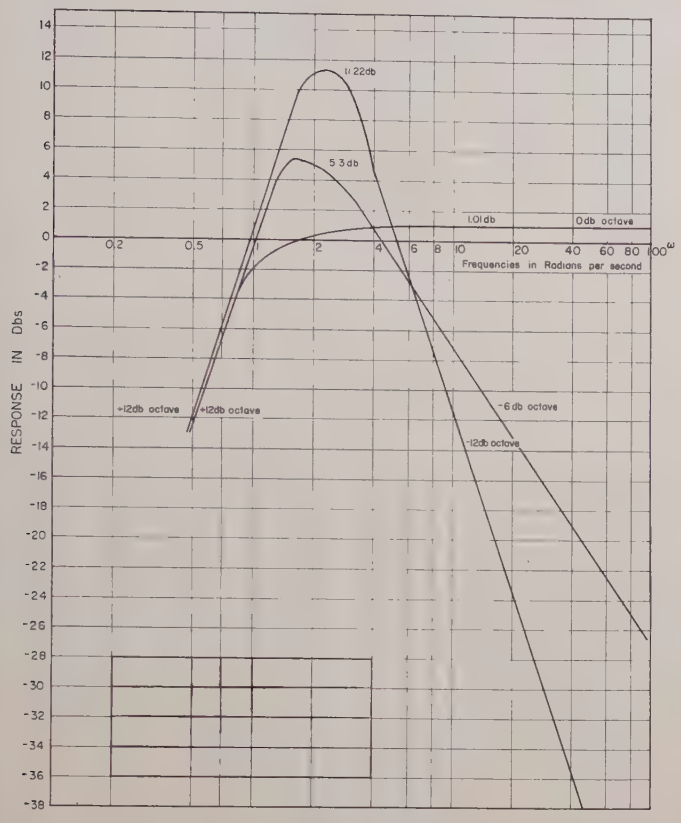


Fig. 2—Frequency response of plant-input function $M(s) = I(s)/R(s)$

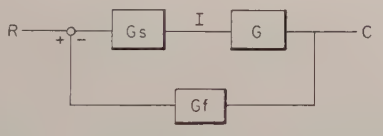


Fig. 3—System configuration for the illustrated problem.

Conclusion

The method of implementing the optimum control system for minimum high-frequency power input to the plant has been presented. The choice of the weighting function for the plant-input function has been discussed and illustrated. The system thus obtained is certain not to excite the high-frequency resonant modes of the plant.

It should be pointed out that the applicability of the proposed method lies on the condition that the frequencies of the plant resonant modes are higher than those of the input signal. In the case where the frequencies of the plant resonant modes are close to the signal frequencies, the exact plant transfer function, including all its resonant modes, should be used in the optimizing procedure.

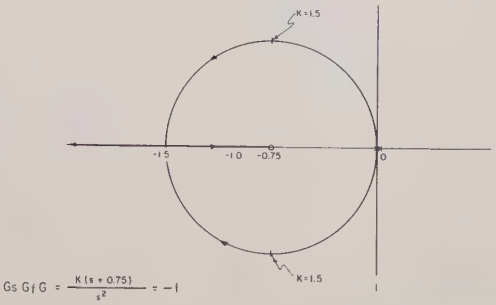


Fig. 4—Root-loci of illustrated problem for $W = 1$.

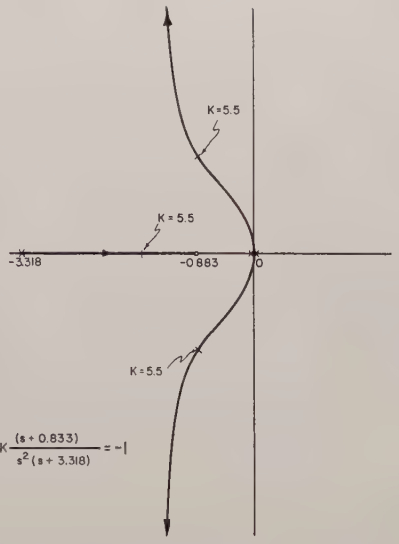


Fig. 5—Root-loci of illustrated problem for $W = s$.

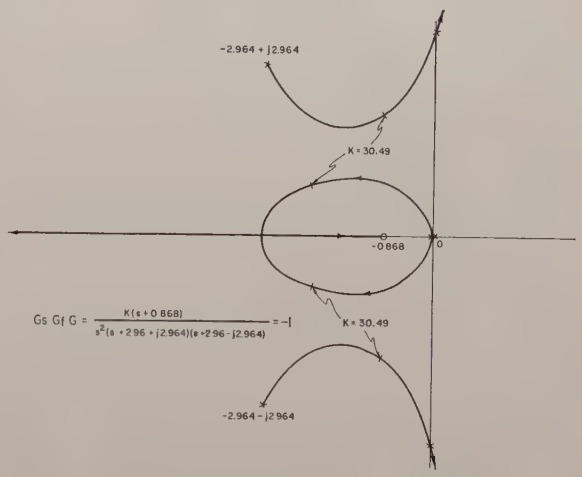


Fig. 6—Root-loci of illustrated problem for $W = s^2$.

Acknowledgment

The author wishes to express his thanks to Dean J. R. Ragazzini and Professor S. S. L. Chang of New York University for their guidance and encouragement.

A Network Theory for Carrier-Suppressed Modulated Systems*

GERALD WEISS†, SENIOR MEMBER, IRE

Summary—**The performance of linear networks in the presence of carrier-suppressed modulation is re-examined in the light of the latest advances in theory. Both analysis and synthesis methods are presented.**

Introduction

QUITE early in the development of feedback control systems it became apparent that the electrical portions of such systems could be considerably simplified if modulated signals were to be used. In particular, the use of carrier-suppressed modulation seemed to lead to appreciable equipment economies. At that time the analysis of a modulated system could be carried out only on an approximate basis, but the performance of the equipment verified the validity of the approximations. Since these early days, carrier frequency (or ac) control systems have been designed and built in complete confidence, despite the absence of an exact theory.

Fifteen years later theory has finally caught up with practice. As a result of the work of a number of authors, it is now possible to construct a unified and reasonably complete theory of linear carrier-frequency networks, covering both analysis and synthesis. This theory is re-presented in this paper. Included are several novel features: 1) a method of obtaining the zeros of envelope functions, using root locus techniques, and the application of the method to the commonly used lead networks; 2) a derivation of the low-pass band-pass transformation as a solution to the synthesis problem, using a logical approximation method; 3) a derivation of the commonly used RC network functions by the same approximation; 4) a numerical evaluation of the approximation, demonstrating the inherent "wide-band-ness" of these so-called "narrow-band" networks.

Carrier-Suppressed Modulation

The conventional ac control system uses carrier-suppressed modulation with a sinusoidal carrier. The unmodulated signal $e(t)$ becomes the modulated signal $e(t) \cos \omega_c t$, where ω_c is the carrier frequency. If such a modulated signal is applied to a linear network, the output is of the form

$$e_0(t) = e_1(t) \cos \omega_c t + e_2(t) \sin \omega_c t, \tag{1}$$

that is, it is made up of *two* modulated voltages. Now if these two voltages are in turn applied to a linear network, the output is of the same form. Hence one can say that the pair of voltages in (1) forms a complete set, or closed set, with respect to carrier-suppressed modulation by a sine wave. The resolution into sine and cosine components is quite arbitrary, and one could just as readily write

$$e_0(t) = e_1'(t) \cos (\omega_c t - 10°) - e_2'(t) \cos (\omega_c t + 59\tfrac{1}{2}°). \tag{2}$$

The signal resolution of (1) is more practical because it is orthogonal with respect to the operation of demodulation. For example, a carrier-suppressed demodulator such as a two-phase induction motor or a ring demodulator can be aligned to "accept" the component $e_1(t)$ and "reject" the component $e_2(t)$. Hence one may view the signal form of (1) a complete orthogonal set for this type of modulation.

Consider now a linear network with input signal

$$r(t) = r_1(t) \sin \omega_c t + r_2(t) \cos \omega_c t. \tag{3}$$

The output is

$$c(t) = c_1'(t) \sin \omega_c t + c_2'(t) \cos \omega_c t. \tag{4}$$

This is a perfectly adequate way of describing the input and output signals. A more general description, however, is the one proposed by Panzer[1] and shown in Fig. 1. The input-output pair is now

$$r(t) = r_1(t) \sin \omega_c t + r_2(t) \cos \omega_c t,$$
$$c(t) = c_1(t) \sin (\omega_c t + \phi) + c_2(t) \cos (\omega_c t + \phi). \tag{5}$$

Fig. 1.—Envelope network response.

$$r(t) = r_1(t) \sin \omega_c t + r_2(t) \cos \omega_c t,$$
$$c(t) = c_1(t) \sin (\omega_c t + \phi) + c_2(t) \cos (\omega_c t + \phi).$$

* Received by the PGAC, May 13, 1960; revised manuscript received October 12, 1960. This paper is an extract from the dissertation "Carrier Frequency Networks," submitted in partial fulfillment of the requirements for the degree of Doctor of Electrical Engineering at the Polytechnic Inst. of Brooklyn. The investigation was carried out at the Microwave Res. Inst. and was supported by the U. S.A.F., Rome Air Dev. Center, under Contract No. AF-30(602)-1648.

† Dept. of Elec. Engrg., Polytechnic Inst. of Brooklyn, Brooklyn, N. Y.

[1] M. Panzer, "Envelope transfer function analysis in a-c servosystems," *AIEE Trans.*, vol. 75, pt. 2, pp. 274–279; November, 1956.

Both $r(t)$ and $c(t)$ are expressed in orthogonal form, but there is now an arbitrary phase angle ϕ between the input and output signal resolutions. This form of signal description is useful in practical ac servos where the demodulator phase alignment is not perfect. For example, in a simple synchro servo, $r_1(t)$ may be the error signal generated by a pair of synchros, and $r_2(t)$ may represent quadrature terms in the synchro output. The servomotor main field may be misaligned with respect to the synchro output $r_1(t)$ by the angle ϕ, possibly because of the phase shift through the synchros. The performance of a network inserted between synchro and motor is then best described by the signal resolution shown in Fig. 1. This particular resolution permits evaluation of the network performance in the presence of both quadrature and phase misalignment. The network envelope performance is thus deliberately made a function of the arbitrary carrier phase resolution angle ϕ.

A more compact signal description is attained by replacing the sinusoids by complex exponentials, as proposed by Chang.[2,3] Eq. (3) becomes

$$r(t) = \mathrm{Im}_i\,[r_1(t) + ir_2(t)]e^{i\omega_c t} \tag{6}$$

or, in the usual short-hand ac notation

$$r(t) = r_1(t) + ir_2(t). \tag{7}$$

One can also define the Laplace transform

$$R_1(p) = \mathfrak{L}[r_1(t)] \qquad p = \sigma_m + j\omega_m \tag{8}$$

and similarly for $R_2(p)$, $C_1(p)$, and $C_2(p)$. Every carrier-suppressed signal has two envelopes, and each can be expressed as a time function or as a frequency function. In the complex frequency domain the two envelopes can be combined into the function

$$R_{\mathrm{env}}(p) = R_1(p) + iR_2(p). \tag{9}$$

This notation was suggested by Panzer[1] and he calls $R_{\mathrm{env}}(p)$ the "entire" envelope function. Note that p is a complex variable in j, hence, $R_{\mathrm{env}}(p)$ is a doubly-complex function in both i and j.

In the discussion to follow, it is assumed that all the envelopes are bandwidth limited to ω_c, the carrier frequency. In other words, the envelope $e(t)$ may have a spectrum as shown in Fig. 2(a). The corresponding spectrum of the modulated signal $e(t)\cos\omega_c t$ is then shown in Fig. 2(b). If the envelope bandwidth is limited as shown, there is no overlap in the sideband spectra and the envelope is recoverable by demodulation.

The question immediately arises as to what happens when the envelope is a step or a ramp, signals with

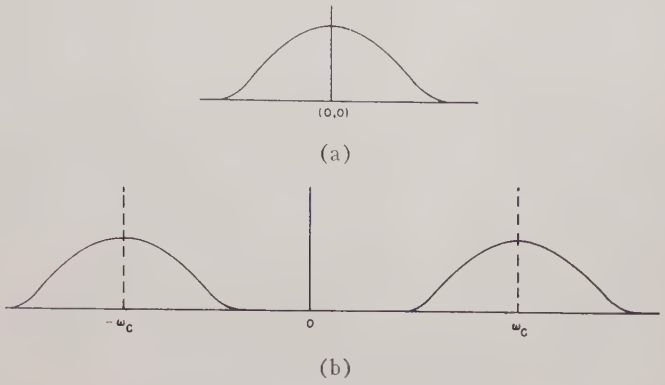


Fig. 2.—Spectrum of modulated signal. (a) Envelope spectrum. (b) Signal spectrum.

infinite bandwidth. Experimental investigations of ac systems do not reveal any difficulties, and perhaps this question is of little importance in today's servo practice. The theoretical problem is puzzling, however, and similar problems have arisen with other types of modulation, for example, in sampled data systems.

In any case, it has been pointed out by Chang[2] that the signal description of (1) is not unique when the envelopes are allowed to have spectral components at the carrier frequency. Hence the assumption of bandwidth limitation is desirable at this point.

Envelope Transfer Functions

The concept of the transfer function is basic to all current control system analysis methods. The transfer function of an element or network for unmodulated signals is simply the ratio of the Laplace transform of the output to the Laplace transform of the input. The transfer function of a demodulator such as the servo induction motor is the ratio of the Laplace transforms of the output shaft angle to the *envelope* of the control field voltage. The transfer function of a modulator such as the ac induction tachometer is the ratio of the Laplace transforms of the *envelope* of the output voltage to the shaft angle. Hence the familiar transfer functions of these electromechanical devices are really a species of envelope transfer function.

It is desirable to employ the same scheme for the transfer function of an electrical network used with modulated signals. The significant quantities are the envelopes, hence the network characteristic of interest is the *envelope* transfer function, the ratio of the Laplace transforms of the envelopes. An immediate difficulty arises, however, because both the input and output signals have two envelopes, not one. Hence any of the four ratios

$$\frac{C_1(p)}{R_1(p)} \qquad \frac{C_1(p)}{R_2(p)} \qquad \frac{C_2(p)}{R_1(p)} \qquad \frac{C_2(p)}{R_2(p)}$$

would appear to qualify as envelope transfer functions.

[2] S. S. L. Chang, "Transient analysis of a-c servomechanisms," *AIEE Trans.*, vol. 74, pt. 2, pp. 30–37; March, 1955.

[3] S. S. L. Chang, "On the separability of Laplace transform variable and its applications in carrier systems," IRE Convention Record, pt 2, pp. 27–34; 1955.

This conceptual difficulty is immediately overcome by use of the doubly-complex notation of (9), and one can define an envelope transfer function

$$G_{\text{env}}(p) = G_D(p) + iG_Q(p) = \frac{C_{\text{env}}(p)}{R_{\text{env}}(p)}. \tag{10}$$

Inserting the definitions of $C_{\text{env}}(p)$ and $R_{\text{env}}(p)$, one obtains

$$C_1(p) + iC_2(p) = [G_D(p) + G_Q(p)][R_1(p) + iR_2(p)] \tag{11}$$

which can be rewritten

$$\begin{aligned} C_1(p) &= G_D(p)R_1(p) - G_Q(p)R_2(p) \\ C_2(p) &= G_Q(p)R_1(p) + G_D(p)R_2(p) \end{aligned} \tag{12}$$

or in the form of a matrix equation

$$\begin{vmatrix} C_1(p) \\ C_2(p) \end{vmatrix} = \begin{vmatrix} G_D(p) & -G_Q(p) \\ G_Q(p) & G_D(p) \end{vmatrix} \begin{vmatrix} R_1(p) \\ R_2(p) \end{vmatrix}. \tag{13}$$

The doubly-complex quantities are thus most readily viewed as skew symmetrical matrices. Subscripts 1 and 2 signify envelopes for the sine and cosine components of the carrier, respectively. The envelope transfer function component $G_D(p)$, the *direct* envelope transfer function, is seen to relate the sine component of the output to the sine component of the input, and the cosine component of the output to the cosine component of the input. The envelope transfer function component $G_Q(p)$, the *quadrature* envelope transfer function, relates the sine component of the output to the cosine component of the input, and vice versa.

The response of a linear network to carrier-suppressed modulated signals is thus characterized by *two* envelope transfer functions $G_D(p)$ and $G_Q(p)$, which can be combined into one doubly complex function $G_{\text{env}}(p)$. Envelope function theory concerns itself with two topics, analysis and synthesis. The analysis problem is: given the network transfer function in the usual sense, what are the envelope transfer functions? The synthesis problem is the converse, but in its usual practical application the problem is: given the *direct* envelope transfer function of a network, what is the transfer function of the network in the usual sense?

The arguments here presented apply not only to transfer functions, but to all network functions, driving point or transfer, impedances, admittances, voltage ratios, and current ratios.

The Analysis Problem

The problem of determining the envelope function of a given network has been considered by a number of authors. Following the pioneering work of Sobczyk,[4] the first to define the envelope transfer functions was Biernson;[5] he derived the relationship between the envelope functions and the usual network function based on a sinusoidal envelope. Chang[2,3] and Candy[6] introduced the concept of the doubly-complex variable. Tou,[7] Panzer[1] and Levenstein[8] presented derivations for arbitrary envelopes. The derivation of Panzer is the most rigorous and complete; given a network with transfer function $H(s)$, the envelope function is

$$H_{\text{env}}(p) = H_{\text{env}}(p, \phi, \omega_c) = H(p + i\omega_c)e^{-i\phi}. \tag{14}$$

The envelope function is thus seen to be dependent on three phenomena: the frequency dependence of the original network as expressed by the complex frequency p, the arbitrary carrier phase resolution angle ϕ, and the carrier frequency ω_c. The two components are given by

$$\left.\begin{aligned} H_D(p, \phi, \omega_c) &= \text{Re}_i\,[H(p + i\omega_c)e^{-i\phi}] \\ H_Q(p, \phi, \omega_c) &= \text{Im}_i\,[H(p + i\omega_c)e^{-i\phi}] \end{aligned}\right\}. \tag{15}$$

It is convenient to define the functions for $\phi = 0$

$$\left.\begin{aligned} H_{D0}(p, \omega_c) &= H_D(p, 0, \omega_c) \\ &= \tfrac{1}{2}[H(p + i\omega_c) + H(p - i\omega_c)] \\ H_{Q0}(p, \omega_c) &= H_Q(p, 0, \omega_c) \\ &= \frac{1}{2i}[H(p + i\omega_c) - H(p - i\omega_c)] \end{aligned}\right\} \tag{16}$$

where

$$\left.\begin{aligned} H_D &= H_{D0}\cos\phi + H_{Q0}\sin\phi \\ H_Q &= -H_{D0}\sin\phi + H_{Q0}\cos\phi \end{aligned}\right\}. \tag{17}$$

Finally, at real frequencies $p = j\omega_m$, one obtains Biernson's formulas[5]

$$\left.\begin{aligned} H_{D0}(j\omega_m, \omega_c) &= \frac{1}{2}\{H[j(\omega_c + \omega_m)] \\ &\quad + H^*[j(\omega_c - \omega_m)]\} \\ H_{Q0}(j\omega_m, \omega_c) &= \frac{1}{2j}\{H[j(\omega_c + \omega_m)] \\ &\quad - H^*[j(\omega_c - \omega_m)]\} \end{aligned}\right\} \tag{18}$$

where * designates the complex conjugate. It should be pointed out that the various authors do not agree in

[4] A. Sobczyk, "Stabilization of carrier-frequency servomechanisms. Part I: Gain-phase margin diagrams of controller characteristics," *J. Franklin Inst.*, vol. 246, pp. 21–43; July, 1948.

[5] G. A. Bjornson, "Network synthesis by graphical methods for a-c servomechanisms," *AIEE Trans.*, vol. 70, pp. 619–625; 1951.

[6] C. J. N. Candy, "A vector method for amplitude-modulated signals," *Proc. IEE*, vol. 103, pt. B, pp. 410–418; 1956.

[7] J. Tou, "Analysis of feedback control systems containing carrier frequency circuits," *Proc. Natl. Electronic Conf.*, vol. 11, pp. 1012–1016; 1955.

[8] H. Levenstein, "On the design of ac networks for servo compensation," IRE Trans. on Automatic Control, vol. AC-2, pp. 39–55; February, 1957.

their definition of the sign of H_Q. The presentation here follows that of Panzer;[1] the convention used by the other authors would require a reversal of sign in (10).

Eq. (18) can be used to obtain the envelope function of any device for which the usual gain and phase vs frequency curves are known. Eq. (15) and (16) are useful when an analysis is to be made in terms of poles and zeros. The relationship between $H(s)$, $H_D(p)$, and $H_Q(p)$ becomes clearer when the functions are expressed in terms of their poles and the principal part of their Laurent series.[8] Table I indicates this relationship for a few special cases such as simple poles, real poles, complex poles, etc. It is seen that each pole of the network function $H(s)$ converts to a pair of complex poles of the envelope functions, displaced from the original poles by $-j\omega_c$. Furthermore, there is a definite relationship between all three functions: if either one is given, the other two are completely determined.

The location of the zeros of the envelope functions can be obtained by a conformal mapping technique analogous to the root locus method. This technique was first applied to a related problem by Smith.[9]

[9] O. J. M. Smith, "Feedback Control Systems," McGraw-Hill Book Co., Inc., New York, N. Y., ch. 18, p. 607ff; 1958.

Starting out with (16) for the case of $\phi=0$, it is desired to solve

$$\left.\begin{aligned} 2H_{D0}(p) &= H(p+j\omega_c) + H(p-j\omega_c) = 0 \\ 2jH_{Q0}(p) &= H(p+j\omega_c) - H(p-j\omega_c) = 0 \end{aligned}\right\}. \tag{19}$$

One can write

$$H(p+j\omega_c) + KH(p-j\omega_c) = 0. \tag{20}$$

Assuming K to be also a function of p, $K(p)$, the procedure is to map the real axis of the K-plane onto the p-plane. At $K=1$ the locus gives the zeros of H_{D0}, and at $K=-1$ the zeros of H_{Q0}. The complete locus will hereafter be referred to as the envelope locus of the function. To facilitate the mapping, the functions are expressed as a ratio of polynomials $N(s)/D(s)$. The mapping equation is then

$$N(p+j\omega_c)D(p-j\omega_c)+K(p)N(p-j\omega_c)D(p+j\omega_c)=0 \tag{21}$$

or, in short-hand notation

$$(N^+)(D^-) + K(N^-)(D^+) = 0, \tag{22}$$

where the roots of all the D and N functions are known. Denote the roots of (N^+) and (D^-) as "poles," and the roots of (N^-) and (D^+) as "zeros." The roots of (24) can then be obtained by standard root locus technique. Posi-

TABLE I

Pole and Residue Relationships

	$e^{-j\phi}H(s)$	$H_D(p)$	$H_Q(p)$
1	$e^{-j\phi}A_0$	$A_0\cos\phi$	$-A_0\sin\phi$
2	$e^{-j\phi}A_1 s$	$A_1(p\cos\phi+\omega_c\sin\phi)$	$A_1(-p\sin\phi+\omega_c\cos\phi)$
3	$e^{-j\phi}\dfrac{A}{(s+a)^n}$	$\dfrac{\frac{1}{2}Ae^{j\phi}}{(p+a-j\omega_c)^n}+\dfrac{\frac{1}{2}Ae^{-j\phi}}{(p+a+j\omega_c)^n}$	$\dfrac{\frac{j}{2}Ae^{j\phi}}{(p+a-j\omega_c)^n}-\dfrac{\frac{j}{2}Ae^{-j\phi}}{(p+a+j\omega_c)^n}$
4	A_0	A_0	0
5	$A_1 s$	$A_1 p$	$A_1\omega_c$
6	$\dfrac{A}{s+a}$	$\dfrac{\frac{1}{2}A}{p+a-j\omega_c}+\dfrac{\frac{1}{2}A}{p+a+j\omega_c}$	$\dfrac{\frac{j}{2}A}{p+a-j\omega_c}-\dfrac{\frac{j}{2}A}{p+a+j\omega_c}$
7	$\dfrac{A-jB}{s+a+jb}+\dfrac{A+jB}{s+a-jb}$	$\dfrac{\frac{1}{2}(A+jB)}{p+a-j(b-\omega_c)}+\dfrac{\frac{1}{2}(A-jB)}{p+a+j(b-\omega_c)}$ $+\dfrac{\frac{1}{2}(A+jB)}{p+a-j(b+\omega_c)}+\dfrac{\frac{1}{2}(A-jB)}{p+a+j(b+\omega_c)}$	$-\dfrac{\frac{1}{2}j(A+jB)}{p+a-j(b-\omega_c)}+\dfrac{\frac{1}{2}j(A-jB)}{p+a+j(b-\omega_c)}$ $+\dfrac{\frac{1}{2}j(A+jB)}{p+a-j(b+\omega_c)}-\dfrac{\frac{1}{2}j(A-jB)}{p+a+j(b+\omega_c)}$
8	$\dfrac{A-jB}{s+a+j\omega_c}+\dfrac{A+jB}{s+a-j\omega_c}$	$\dfrac{A}{p+a}+\dfrac{\frac{1}{2}(A+jB)}{p+a-j2\omega_c}+\dfrac{\frac{1}{2}(A-jB)}{p+a+j2\omega_c}$	$\dfrac{B}{p+a}+\dfrac{\frac{1}{2}j(A+jB)}{p+a-j2\omega_c}-\dfrac{\frac{1}{2}j(A-jB)}{p+a+j2\omega_c}$

tive real K's map into the 180° locus, and negative real K's map into the 0° locus. The $K=\pm 1$ points are obtained by measuring the ratio of pole and zero distances. This construction may not even be necessary because it is apparent that $|K|=1$ along the real axis.

A simple example of the technique is shown in Fig. 3 for an RC coupling network. In case (b) the zeros of G_{D0} are the real axis intercepts, in case (c) they are located on the vertical axis of symmetry for the zero-pole constellation. The zeros of G_{Q0} are at infinity; there are two branches going to infinity, therefore G_{Q0} has a double zero. Of course, the zeros of this simple network could easily have been obtained algebraically. The utility of the method becomes apparent when the network is at least a biquadratic.

Several features of this type of conformal mapping may be worth noting. The critical points designated as "poles" and "zeros" do not occur in complex conjugate pairs, but for each "pole" there is a complex conjugate "zero." Since the relationship between the (N^+) and (N^-) functions is

$$N^*(p+j\omega_c) = N(p^*-j\omega_c) \tag{23}$$

and similarly for D, one can derive from (21)

$$K^*(p) = \frac{1}{K(p^*)}. \tag{24}$$

This means that if a real value of K, say K_1, maps into p, another real value $K_2=1/K_1$, maps into p^*. Hence the complete envelope locus is symmetrical about the real $-p$ axis, but the mapping of the individual K-point is not, except at $K=\pm 1$.

The polynomials $(N^+)(D^-)$ and $(N^-)(D^+)$ are of the same degree and the coefficients of the leading term are identical. Hence at least one of the $K=-1$ points is at infinity,[10] in addition to any degenerate locus at infinity. An envelope locus with a single $K=-1$ point at infinity pictorially resembles the conventional root locus of a feedback system of asymptotic order 2, because there is one half-branch going to and a second half-branch leaving the $K=-1$ point. The functions shown here have two infinite H_{Q0} zeros, so that their envelope loci resemble the root locus of feedback systems of asymptotic order 4.

Now consider the general case of $\phi \neq 0$. From (17) one obtains

$$(N^+)(D^-) + Ke^{j2\phi}(N^-)(D^+) = 0. \tag{25}$$

The zeros of (25) are obtained by mapping the unit circle of the K-plane onto the p-plane, using (22) as the mapping equation. Along this circle $K(p)$ is equal to $e^{j\theta(p)}$, so that (22) becomes

$$(N^+)(D^-) + e^{j\theta(p)}(N^-)(D^+) = 0. \tag{26}$$

[10] H. Ur, "Root locus properties and sensitivity relations in control systems," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 57–65; January, 1960.

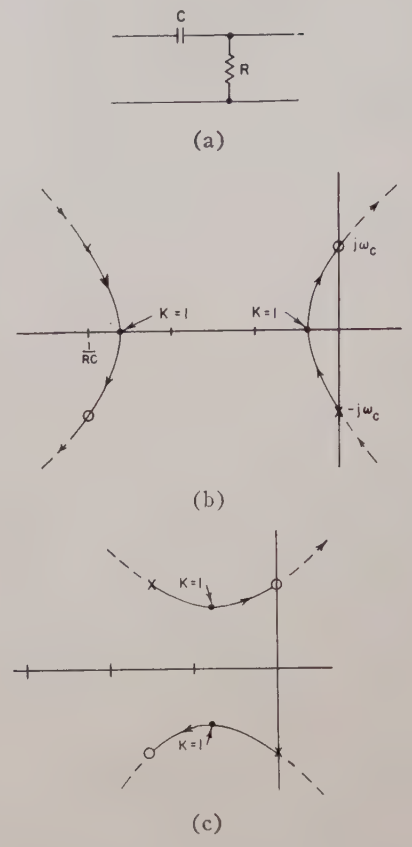


Fig. 3.—Root locus of coupling network.

(a) RC coupling network.

$$G(s) = \frac{s}{s+\frac{1}{RC}},$$

$$G(p \pm j\omega_c) = \frac{p \pm j\omega_c}{\left(p+\frac{1}{RC}\right) \pm j\omega_c}.$$

(b) Locus for

$$\frac{1}{RC} = 3\omega_c.$$

(c) Locus for

$$\frac{1}{RC} = \frac{3}{2}\omega_c.$$

For (b) and (c), arrows show direction of increasing K;
——— $K>0$,
- - - - - - - $K<0$.

The zeros of $H_D(p, \phi)$ are located at the mapping of $\theta=2\phi$; the zeros of $H_Q(p, \phi)$ are located at the mapping of $\theta=2\phi-\pi$. Fig. 4 is a simple example of such a locus, for the same network as in Fig. 3(b). The unit circle mapping is, of course, orthogonal to the envelope locus mapping, and exhibits the same asymptotic order. The entire real-p axis must be a portion of the unit circle mapping. Each point on the unit circle of the K-plane maps into pairs of conjugate complex (or real) points in the p-plane, because

$$\theta(p^*) = \theta(p). \tag{27}$$

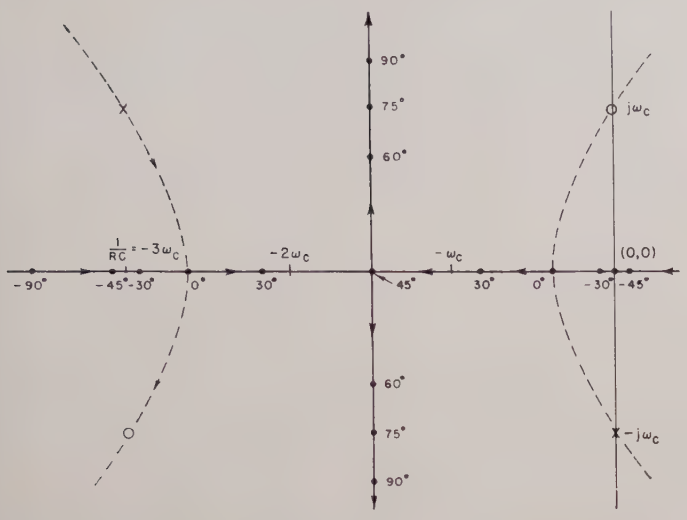


Fig. 4—Unit circle mapping for coupling network.

$$G(s) = \frac{s}{s + \frac{1}{RC}} \left(\text{Plotted for } \frac{1}{RC} = 3\omega_c \right).$$

——— locus of $|K| = 1$, direction increasing θ.
- - - - - - - locus of K real, direction increasing K.

Zeros of AC Lead Network

The commonly used ac lead networks, bridged-T, parallel-T, and resonant damper, are characterized by the biquadratic transfer function

$$G(s) = \frac{s^2 + 2as + \omega_c^2}{s^2 + 2bs + \omega_c^2} \qquad a < b, \qquad a < \omega_c. \tag{28}$$

This function is obtained through the low-pass band-pass transformation of the ordinary lead compensation function

$$G(p) = \frac{p + a}{p + b} \qquad a < b. \tag{29}$$

In other words, the direct envelope function G_{D0} of this network is supposed to approximate the bilinear function of (29). In particular, it is intended that G_{D0} have a dominant zero near $-a$. In actuality G_{D0} is represented by a biquartic function, the poles of which are easily determined by using Table I. The numerator quartic of G_{D0} is, however, not factorable; an approximate location of the zeros can then be obtained by the mapping procedure outlined previously. The poles and zeros of (28) are located, respectively, at

$$\begin{aligned} s &= -b \pm \sqrt{b^2 - \omega_c^2} \\ s &= -a \pm j\sqrt{\omega_c^2 - a^2}. \end{aligned} \tag{30}$$

One now forms the function $G(p \pm j\omega_c)$ and sets up the equation

$$G(p + j\omega_c) + K(p)G(p - j\omega_c) = 0. \tag{31}$$

For the purposes of the conformal mapping, the "poles" are at

$$\left.\begin{aligned} p &= -a - j[\omega_c \pm \sqrt{\omega_c^2 - a^2}] \\ p &= -b \pm \sqrt{b^2 - \omega_c^2} + j\omega_c \end{aligned}\right\}. \tag{32}$$

The "zeros" are at

$$\left.\begin{aligned} p &= -a + j[\omega_c \pm \sqrt{\omega_c^2 - a^2}] \\ p &= -b \pm \sqrt{b^2 - \omega_c^2} - j\omega_c \end{aligned}\right\}. \tag{33}$$

The loci are shown in Figs. 5 to 7, for the three cases $b < \omega_c$, $b = \omega_c$, and $b > \omega_c$.

The zeros of G_Q are at $K = -1$, and there are double zeros at the origin and at infinity. The zeros of G_D are at $K = 1$. In each case there is

1) One real axis zero slightly to the left of $p = -a$.
2) One real axis zero near $\mathrm{Re}[-b - \sqrt{b^2 - \omega_c^2}]$.
3) One pair of conjugate complex zeros, the real part of which is somewhere between $-a$ and $\mathrm{Re}[-b + \sqrt{b^2 - \omega_c^2}]$, and the imaginary part of which is somewhere between ω_c and $2\omega_c$.

The zero-pole constellations of G_D are shown in Fig. 8 (page 61). In every case, there is one dominant critical point, the zero near $p = -a$. The distance from the origin of the other zeros is never less than b or ω_c, whichever is smaller.

It remains to investigate the effect of shifting the carrier component resolution angle ϕ. The unit circle locus shown in Fig. 9 was constructed for $b > \omega_c$. For the other cases, $b \leq \omega_c$, the unit circle locus is similar in shape. It is seen that considerable variations in ϕ hardly affect the predominant zero of $G_D(p, \phi)$, which remains near $p = -a$ for $45° < \phi < 45°$. Within this large range the other real zero is still rather far away. The complex zeros move into the right half-plane when ϕ becomes somewhat negative, and have moved relatively close to the origin by $\phi = 45°$.

The zeros of $G_Q(p, \phi)$, on the other hand, are quite sensitive to small variations in ϕ. The zeros at the origin move rapidly right and left along the real axis for $\phi > 0$, and along the small circle for $\phi < 0$. The zeros at infinity come in as real zeros for $\phi > 0$, and as complex conjugate zeros for $\phi < 0$, but they remain at a considerable distance from the origin. All zeros are real for $\phi > 0$, conjugate complex for $\phi < 0$.

The Synthesis Problem

From the foregoing it is clear that an envelope function, to be realizable as a linear network, must have one of the pole-residue patterns listed in Table I. If it does, the network function $H(s)$ can be written down immediately from the partial fraction expansion of the particular $H(p)$. Synthesis of the network then becomes a matter of standard technique.

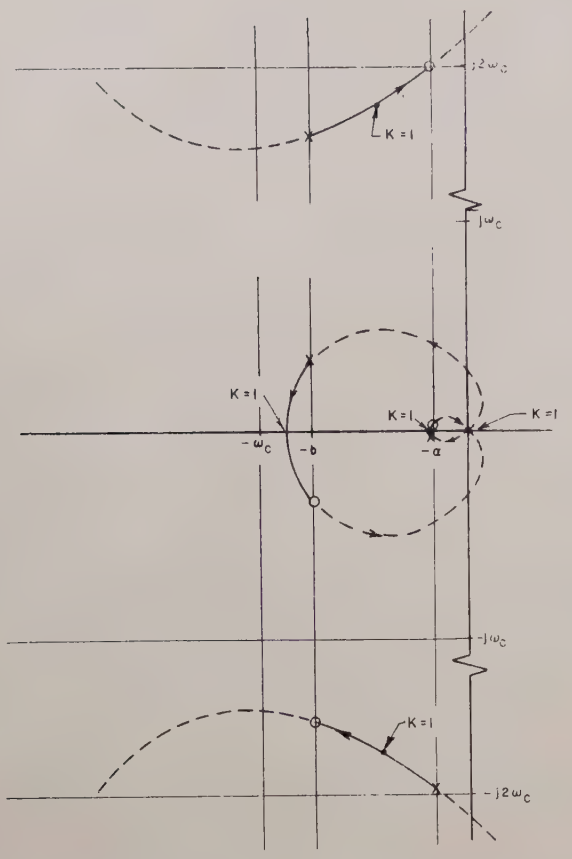


Fig. 5—Locus of envelope function zeros for complex poles.
——— $K>0$.
- - - - - - - $K<0$.
— → — direction of increasing K.

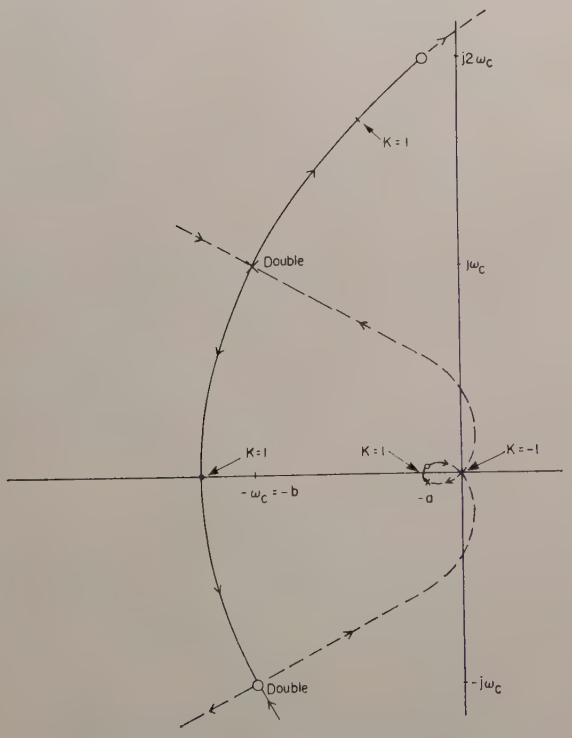


Fig. 6—Locus of envelope function zeros for double pole.
——— $K>0$.
- - - - - - - $K<0$.
— → — direction of increasing K.

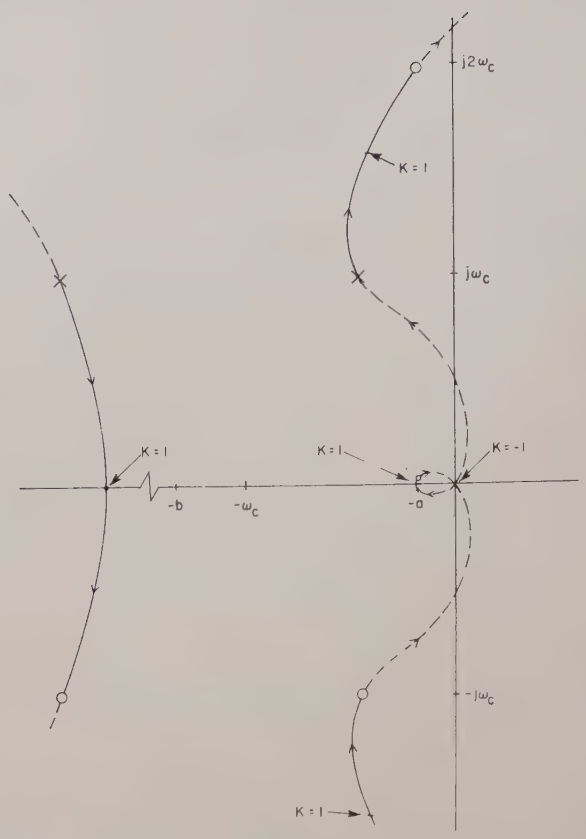


Fig. 7—Locus of envelope function zeros for real poles.
→→→→ direction of increasing K.
——— positive K.
- - - - - - - negative K.

In most practical problems, however, both $H_D(p)$ and $H_Q(p)$ are specified independently, and they are in general incompatible: that is, they cannot be realized by the same network. A common requirement on H_Q is for it to be identically zero, which can be realized only by the trivial network $H(s)=A_0$. In any case, even if the desired $H_D(p)$ and $H_Q(p)$ are compatible it is unlikely that they are realizable because the allowable pole-residue patterns shown in Table I are very specialized. Normally $H_D(p)$ will be given analytically, or $H_D(j\omega_m)$ might be given graphically, and must be approximated by a polynomial. In either case, the most commonly encountered form of specification, the first order polynomial, is not realizable; the next most common one, the second order polynomial, is allowed only in a very limited way. It can therefore be stated that a realizable set of conditions on both $H_D(p)$ and $H_Q(p)$ is never encountered in practice. Actual synthesis problems therefore can never be solved except by some kind of approximation.

A somewhat different approach to the synthesis procedure has been described by Hellerman.[11] He requires the availability of a quadrature input signal, *i.e.*, a signal with the same modulation as the original signal, but on the quadrature carrier. Each of these signals is then

[11] H. Hellerman, "Transfer functions for amplitude-modulated signals," *AIEE Trans.*, vol. 74, pt. 3, pp. 729–731; January, 1956.

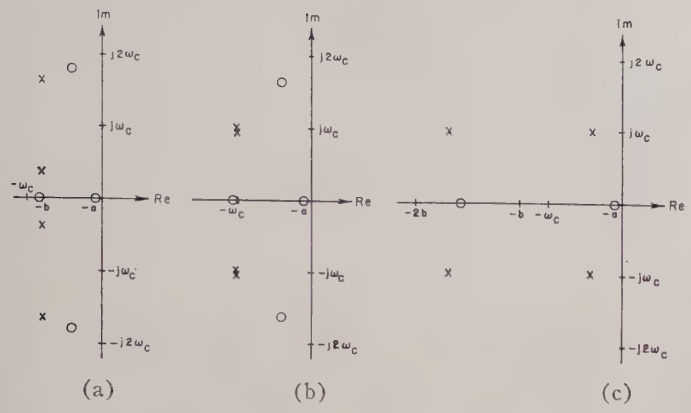


Fig. 8—Zero-pole constellations for direct envelope function of lead network

$$G(s) = \frac{s^3 + 2as + \omega_c^2}{s^3 + 2bs + \omega_c^2}.$$

(a) $b < \omega_c$. (b) $b = \omega_c$. (c) $b > \omega_c$.

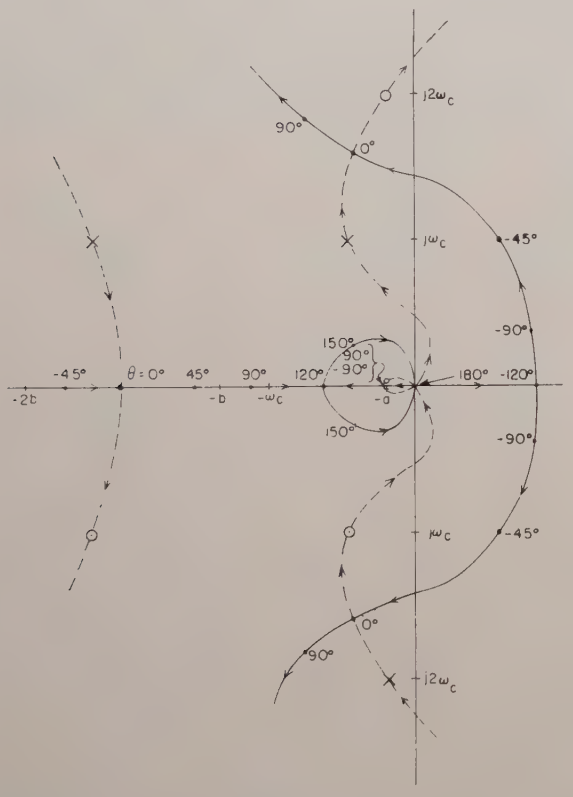


Fig. 9—Locus of envelope function zeros with carrier resolution angle ϕ.
Zeros of $G_p(p, \phi)$ at $\theta = 2\phi$.
Zeros of $G_Q(p, \phi)$ at $\theta = 2\phi - \pi$.
——→—— locus of $|K_e^{j\theta}| = 1$, direction of increasing θ.
——→—— locus of K real, direction of increasing K.

passed through separate linear networks $H_D(s)$ and $H_Q(s)$, and the responses are added to form a single modulated output signal with zero quadrature. The relationship between the network functions $H_D(s)$ and $H_Q(s)$ and the specified envelope function $H_D(p)$ is shown to be[11]

$$\left.\begin{aligned} H_D(s) &= \operatorname{Re} H_D(p - j\omega_c) \\ H_Q(s) &= \operatorname{Im} H_D(p - j\omega_c) \end{aligned}\right\}. \tag{34}$$

Hellerman proves that both $H_D(s)$ and $H_Q(s)$ are realizable, provided that $H_D(p)$ is realizable as a "low-pass prototype" network. Again, from a practical point of view, such a quadrature signal is not available. The use of just a single network $H_D(s)$ then again constitutes some kind of approximation.

Historically, all attempts at a logical, direct, synthesis procedure have faltered at this point. A possible reason for this failure is that approximate solutions to the synthesis problem, derived more or less intuitively, appear to work satisfactorily. An excellent review of this approach to envelope network synthesis is presented by Blanton.[12] In order to obtain an identically zero quadrature response, the network $H(s)$ must have the property

$$H[j(\omega_c + \omega_m)] = H^*[j(\omega_c - \omega_m)]. \tag{35}$$

In other words, the amplitude response must have symmetry about the carrier frequency, while the phase response must have skew-symmetry about its value at the carrier frequency. The ideal network function then has a frequency response like that of a physically realizable network but shifted along the frequency axis by an amount ω_c; such a function is not realizable except for the trivial $\omega_c = 0$. There does exist a realizable frequency shift or transformation, namely the low-pass band-pass transformation.[13] If $H_D(p)$ is realizable as "a low-pass prototype" network, $H(s)$ is obtained immediately by setting

$$\frac{p}{\omega_c} = \frac{1}{2}\left(\frac{s}{\omega_c} + \frac{\omega_c}{s}\right). \tag{36}$$

Each capacitor in the $H_D(p)$ network transforms into a parallel LC in the $H(s)$ network and each inductor in $H_D(p)$ transforms into a series LC in $H(s)$, with both LC combinations resonant to the carrier frequency. The frequency response curves of these transformed functions, however, do not show the required symmetry unless a logarithmic frequency scale is used, that is, they exhibit geometric rather than arithmetic symmetry about the carrier frequency. The difference between arithmetically and geometrically symmetrical characteristics vanishes as the origin of symmetry is approached, hence the designation "narrow-band approximation" for this design procedure. The networks are also called "narrow-band networks," even though it soon appears that reasonable Q's or RC realizability require wide-band networks, in the sense that relative bandwidth is defined as the ratio of a critical envelope frequency (say the frequency of maximum phase lead) to the carrier frequency.[14]

[12] H. E. Blanton, "Carrier compensation for servomechanisms," *J. Franklin Inst.*, vol. 250, pp. 391–407, 525–542; 1950.
[13] P. R. Aigran, B. R. Teare and E. M. Williams, "Generalized theory of the band-pass low-pass analogy," PROC. IRE, vol. 37, pp. 1152–1155; October, 1949.
[14] G. Weiss, "Carrier Frequency Networks," Microwave Res. Inst., Polytechnic Inst. of Brooklyn, Brooklyn, N. Y., Rept. No. R-701-58, PIB-629; January, 1959.

It is only recently that Levenstein[8] has shown that the synthesis of envelope networks can be treated in a fashion analogous to the approximation problem of classical network synthesis.

The Approximation Problem

A direct and logical attack on the approximation problem is to approximate the function to be synthesized by one or more of the realizable functions given in Table I. This involves several undetermined coefficients which can be selected in such a way as to optimize the approximation.[8]

The steps in the approximation procedure are analogous to those used in classical network synthesis:

1) Approximate the specified $H_D(p)$ by a function having the permissible pole-residue pattern (see Table I).
2) Compute the $H_Q(p)$ corresponding to the selected $H_D(p)$.
3) Optimize the parameters of the selected $H_D(p)$ to minimize the difference between the actual and desired $H_D(p)$ and $H_Q(p)$, based on a preselected "goodness" criterion.
4) Derive $H(s)$ from the optimized $H_D(p)$.
5) Synthesize the network form $H(s)$ by conventional circuit synthesis.

The first step, the selection of a permissible pole-residue pattern, immediately determines the type of network. Approximation of $H_D(p)$ by conjugate complex poles (line 6, Table I) leads to real poles of $H(s)$ which permits synthesis by RC networks. Approximation of $H_D(p)$ by triplets and quadruplets (lines 7 and 8, Table I) leads to complex poles of $H(s)$ and RLC networks.

The "error" criterion or "goodness" criterion chosen to optimize the selected $H_D(p)$ will determine the detailed structure of the network and its parameter values. It will be shown how a particular set of optimization criteria leads to the conventional low-pass band-pass approximation, and at the same time to the currently used RC lead networks. Thus the whole field of envelope networks is unified.

A Low-Pass Approximation

It is desired to synthesize a network function $H(s)$ which approximates a prescribed direct envelope function $H_D(p)$ and a prescribed quadrature envelope function $H_Q(p)=0$. The actual envelope functions will be designated $H_D'(p)$ and $H_Q'(p)$, respectively. An error function $E_D(p)$ is defined

$$E_D(p) = H_D'(p) - H_D(p). \tag{37}$$

The design procedure will be illustrated for the simple pole

$$H_D(p) = \frac{K}{p+b}. \tag{38}$$

Table I indicates that such a function is not realizable, and the first step is to approximate it by a realizable function $H_D'(p)$. One might select $H_D'(p)$ from line 6, as a conjugate pair. Or one might select an approximation in the form of three poles, line 8, as was done by Levenstein.[8] The procedure to be followed will be to select the most general approximation, the quadruplet of poles, line 7. Hence

$$\left.\begin{aligned} H_D'(p) = {} & \frac{\frac{1}{2}(A+jB)}{p+b'-j(D-\omega_c)} + \frac{\frac{1}{2}(A-jB)}{p+b'+j(D-\omega_c)} \\ & + \frac{\frac{1}{2}(A+jB)}{p+b'-j(D+\omega_c)} + \frac{\frac{1}{2}(A-jB)}{p+b'+j(D+\omega_c)} \end{aligned}\right\} \tag{39}$$

where there are four undetermined real coefficients, the pole parameters b' and D, and the residue parameters A and B. From Table I the corresponding quadrature function is

$$\begin{aligned} H_Q'(p) = {} & -\frac{\frac{1}{2}j(A+jB)}{p+b'-j(D-\omega_c)} + \frac{\frac{1}{2}j(A-jB)}{p+b'+j(D-\omega_c)} \\ & + \frac{\frac{1}{2}j(A+jB)}{p+b'-j(D+\omega_c)} - \frac{\frac{1}{2}j(A-jB)}{p+b'+j(D+\omega_c)} \end{aligned} \tag{40}$$

and the corresponding network function is

$$H(s) = \frac{A-jB}{s+b'+jD} + \frac{A+jB}{s+b'-jD}. \tag{41}$$

Simplifying these expressions, one obtains

$$H_D'(p) = 2\frac{\left\{\begin{aligned} & Ap^3 + (3Ab' - BD)p^2 + [A(3b'^2 + D^2 + \omega_c^2) - 2b'BD]p \\ & + [Ab'(b'^2 + D^2 + \omega_c^2) - BD(b'^2 + D - \omega_c^2)] \end{aligned}\right\}}{[(p+b'^2 + (D-\omega_c)^2][(p+b')^2 + (D+\omega_c)^2]} \tag{42}$$

$$H_Q'(p) = \frac{-2\omega_c\{Ap^2 + 2(Ab' - BD)p + [A(b'^2 - D^2 + \omega_c^2) + 2b'BD]\}}{[(p+b')^2 + (D-\omega_c)^2][(p+b')^2 + (D+\omega_c)^2]} \tag{43}$$

$$H(s) = \frac{2[As + (Ab' - BD)]}{s^2 + 2b's + (b' + D^2)}. \tag{44}$$

One now proceeds to select parameters so as to minimize $H_Q'(p)$. In a low-pass system, like the usual servomechanism, it is plausible to minimize $H_Q'(p)$ at $p=0$. The numerator of (42) is a quadratic, hence there exists the possibility of obtaining a double zero at the origin; in other words, both the quadrature function and its derivative are made to vanish at zero modulating frequency. This result is obtained by setting

$$Ab' - BD = 0 \tag{45}$$

$$A(b'^2 - D^2 + \omega_c^2) - 2b'BD = 0. \tag{46}$$

Simultaneous solution of these two equations yields

$$b'^2 + D^2 = \omega_c^2 \tag{47}$$

so that the network is tuned to the carrier frequency. The relation (45) also simplifies the network (44) by eliminating one numerator term

$$H(s) = \frac{2As}{s^2 + 2b's + \omega_c^2}. \tag{48}$$

Since the parameters b', D, A, and B are all real, the relationship (46) implies $\omega_c > b'$. Thus the network function has complex poles and is realizable only by RLC networks.

Consider instead the case $\omega_c < b'$, that is, D and B imaginary. Let $D=jD$, $B=-jB$. This is equivalent to starting out with a quadruplet

$$H_D'(p) = \frac{\frac{1}{2}(A-B)}{p+b'-D-j\omega_c} + \frac{\frac{1}{2}(A-B)}{p+b'-D+j\omega_c} + \frac{\frac{1}{2}(A+B)}{p+b'+D-j\omega_c} + \frac{\frac{1}{2}(A+B)}{p+b'+D+j\omega_c}. \tag{49}$$

In other words, $H_D(p)$ is now approximated by two conjugate complex pairs, see Table I, line 6. The poles of $H(s)$ are now real, and realization by RC networks is possible. From the table the corresponding quadrature function is

$$H_Q'(p) = \frac{\frac{j}{2}(A-B)}{p+b'-D-j\omega_c} - \frac{\frac{j}{2}(A-B)}{p+b'-D+j\omega_c} + \frac{\frac{j}{2}(A+B)}{p+b'+D-j\omega_c} - \frac{\frac{j}{2}(A+B)}{p+b'+D+j\omega_c} \tag{50}$$

and the corresponding network function is

$$H(s) = \frac{A-B}{s+b'-D} + \frac{A+B}{s+b'+D}. \tag{51}$$

These expressions can again be simplified to expressions resembling (42)–(44). Minimizing the quadrature at $p=0$, one obtains

$$Ab' - BD = 0$$

$$A(b'^2 + D^2 + \omega_c^2) - 2Bb'D = 0. \tag{52}$$

So that

$$b'^2 - D^2 = \omega_c^2 \quad (\omega_c < b') \tag{53}$$

and the network is again tuned. For *either* case one obtains

$$H_D(p) = \frac{2A(p^3 - 2b'p^2 + 2\omega_c^2 p + 2b'\omega_c^2)}{p^4 + 4b'p^3 + 4(b'^2+\omega_c^2)p^2 + 8b'\omega_c^2 p + 4b'^2\omega_c^2} \tag{54}$$

$$H_Q'(p) = \frac{-2\omega_c A p^2}{p^4 + 4b'p^3 + 4(b'^2+\omega_c^2)p^2 + 8b'\omega_c^2 p + 4b'^2\omega_c^2} \tag{55}$$

and

$$H(s) = \frac{2As}{s^2 + 2b's + \omega_c^2}. \tag{56}$$

The preceding optimization procedure for minimum quadrature establishes the pole pattern of the functions $H(s)$, $H_D'(p)$, and $H_Q'(p)$. The pole locations are tabulated in Table II for the three cases $b' < \omega$, $b' = \omega$, and $b' > \omega_c$. A straight-forward geometric construction can be set up. The poles of $H(s)$ are located, for $b' \leq \omega_c$, on the circle of radius ω_c with center at the origin; for $b' \geq \omega_c$, these poles are of course on the real axis. The poles of the envelope functions, for $b' \leq \omega_c$, are located on two circles of radius ω_c, with centers at $(0, \pm j\omega_c)$; for $b' \geq \omega_c$, these poles are located on the circle of radius b' and center at $(-b, 0)$, and its intersection with the horizontal lines $s = \pm j\omega_c$. The poles will travel along loci when the relative values of b' and ω_c are changed. Fig. 10 shows the loci of the poles for fixed carrier frequency and varying b'.

Of the four undetermined coefficients, two, namely B and D, have now been selected. To select the remaining coefficients, A and B', the error function $E_D(p)$ is now minimized. Substituting (38) and (54) into (37),

$$E_D(p) = \frac{\left\{\begin{array}{r}(2A-K)p^4 + 2[A(b+2b') - 2b'K]p^3 + 4[A(\omega_c^2 + bb') - K(\omega_c^2 + b'^2)]p^2 \\ + 4\omega_c^2[A(b+b') - 2b'K]p + 4b'\omega_c^2(Ab - b'K)\end{array}\right\}}{(p+b)[p^4 + 4b'p^3 + 4(\omega_c^2 + b'^2)p^2 + 8b'\omega_c^2 p + 4b'^2\omega_c^2]}. \tag{57}$$

Minimizing this function and its first derivative at $p=0$ yields

$$Ab - b'K = 0$$

$$A(b+b') - 2b'K = 0 \tag{58}$$

hence

$$b' = b, \qquad A = K \tag{59}$$

TABLE II

POLE LOCATIONS FOR ENVELOPE NETWORK FUNCTIONS OPTIMIZED FOR MINIMUM QUADRATURE ("TUNED" NETWORKS)

	poles of $H(s)$	poles of $H_D(p)$ and $H_Q(p)$
$b' < \omega_c$	$-b' + j\sqrt{\omega_c^2 - b'^2}$	$-b' + j(\sqrt{\omega_c^2 - b'^2} + \omega_c)$
		$-b' + j(\sqrt{\omega_c^2 - b'^2} - \omega_c)$
	$-b - j\sqrt{\omega_c^2 - b'^2}$	$-b' - j(\sqrt{\omega_c^2 - b'^2} - \omega_c)$
		$-b' - j(\sqrt{\omega_c^2 - b'^2} + \omega_c)$
$b' = \omega_c$	$-\omega_c$ (double)	$-\omega_c + j\omega_c$ (double) $-\omega_c - j\omega_c$ (double)
$b' > \omega_c$	$-b' + \sqrt{b'^2 - \omega_c^2}$	$(-b' + \sqrt{b'^2 - \omega_c^2}) + j\omega_c$
		$(-b' + \sqrt{b'^2 - \omega_c^2}) - j\omega_c$
	$-b' - \sqrt{b'^2 - \omega_c^2}$	$(-b' - \sqrt{b'^2 - \omega_c^2}) + j\omega_c$
		$(-b' - \sqrt{b'^2 - \omega_c^2}) - j\omega_c$

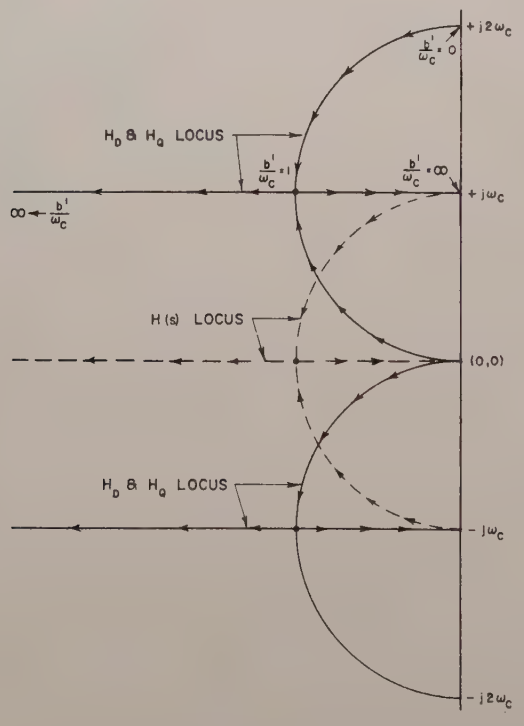


Fig. 10—Pole loci of $H(s)$, $H_D(p)$ and $H_Q(p)$ after quadrature optimization ("tuned" networks).

so that

$$E_D(p) = \frac{K(p+2b)p^3}{(p+b)[p^4+4bp^3+4(\omega_c^2+b^2)p^2+8b\omega_c^2p+4b^2\omega_c^2]}. \quad (60)$$

As a result $E_D(p)$ has a triple pole at the origin, the function itself and its first two derivatives vanish at $p=0$. One also obtains

$$H_D'(p) = \frac{2K(p^3+2bp^2+2\omega_c^2p+2b\omega_c^2)}{p^4+4bp^3+4(\omega_c^2+b^2)p^2+8b\omega_c^2p+4b^2\omega_c^2} \quad (61)$$

$$H_Q'(p) = \frac{-2K\omega_c p^2}{p^4+4bp^3+4(\omega_c^2+b^2)p^2+8b\omega_c^2p+4b^2\omega_c^2} \quad (62)$$

$$H(s) = \frac{2Ks}{s^2+2bs+\omega_c^2} = \frac{K}{\frac{\omega_c}{2}\left(\frac{s}{\omega_c}+\frac{\omega_c}{s}\right)+b}. \quad (63)$$

The relationship between $H_D(p)$, (38), and $H(s)$, (63), is exactly that given by the conventional low-pass band-pass transformation. This correspondence can also be shown for other forms of $H_D(p)$, the linear polynomial, a complex pole, repeated poles, etc.[14]

The approximation method here presented does not immediately yield new networks useful in control engineering practice. It does show, however, how the currently used networks are derived, how they fit the desired specification. In particular, it is seen that the conventional RC and RLC networks are both obtained by the same approximation procedure, and are represented by the same algebraic function; RC networks are obtained for particular pole locations of this function.

Numerical Evaluation

The error functions H_Q' and E_D have multiple zeros at the origin, hence one might expect that the approximation is much better than the term "narrow-band approximation" would lead one to believe. This is indeed the case, as the following numerical example will show. Consider the lead network specification

$$G_D(p) = \frac{p+a}{p+b} = 1 - \frac{b-a}{p+b} \qquad a<b. \quad (64)$$

The values of G_Q' and E_D are obtained by substituting $(a-b)$ for K in (62) and (60); the value of G_D' is obtained further by adding unity to (61). All these functions can be evaluated at real modulating frequencies, $p=j\omega_m$.

It is convenient to introduce the normalizations

$$\left.\begin{aligned} \alpha &= \frac{a}{b} = \text{gain at zero modulating frequency,} \\ \beta &= \frac{\sqrt{ab}}{\omega_c} = \text{relative bandwidth of network,} \\ u &= \frac{\omega_m}{\sqrt{ab}} = \text{normalized modulating frequency.} \end{aligned}\right\} \quad (65)$$

The functions have been evaluated for a low-pass prototype designed for a maximum phase shift of 41.8° at 26.8 cps ($\alpha=0.2$, $a=12$ cps, $b=60$ cps). Two carrier frequencies are considered, 60 cps and 400 cps. The choice of a 400-cps carrier ($\beta=0.0671$) would result in a moderate bandwidth system. The choice of a 60-cps carrier ($\beta=0.447$), on the other hand, results in an extremely wide-band system; in fact, $b=\omega_c$, so that the resultant network will have a double pole on the negative real axis.

The characteristics of the narrow-band network (400-cps carrier frequency) are plotted in Fig. 11. The approximation to the desired direct envelope function is within 0.4 per cent over the range of modulating frequencies of 0–60 cps; it is perfect for all practical purposes. The magnitude of the quadrature term is less than 4 per cent of the direct term. For the wide-band case (60-cps carrier frequency), the functions are plotted on Fig. 12 in polar form, with the relative modulating frequency u as parameter. The approximation is not as good, the error in the magnitude of the direct function becoming as large as 11 per cent; the quadrature function is 40 per cent of the desired direct term for the range of modulating frequencies previously considered. From a practical engineering point of view, the approximation to the direct function is satisfactory at all modulating frequencies. The important criterion here is the phase angle of the direct function. Below $u=0.7$ the phase of G_D' leads the phase of G_D by about 0.1°. The two loci then intersect, and G_D' begins to lag G_D. At $u=1$ this lag is only 0.6°, and it does not become 10° until $u=1.72$. It would appear that this phase lag at high frequencies can easily be taken care of by overspecifying the required phase angle. In any case the example here considered is an extreme one.

The approximation to the quadrature function is not so good. When the ratio $|G_Q/G_D|$ is 0.4, any system error already existing due to the combination of a quadrature signal plus a phase misalignment as small as 10° will be tripled. However, G_Q is zero at zero frequency so that the quadrature error is not significantly increased under steady-state conditions. Under dynamic operating condition the quadrature error due to the network may be considerable, which emphasizes the importance of minimizing both quadrature noise and demodulator phase misalignment.

Another way of looking at the approximation error is to evaluate the functions at $u=1$, the frequency of maximum phase lead for the low-pass prototype. This is done in Fig. 13 for three different values of α. It is seen that the difference in phase lead between the actual network and the low-pass prototype rarely exceeds 5°.

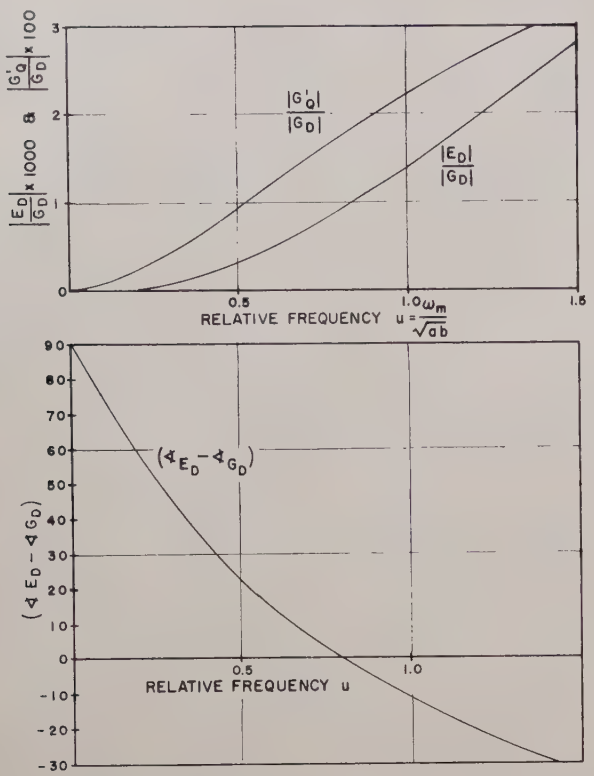


Fig. 11—Approximation function $E_D(j\omega_m)$ for narrow-band case (400 cps carrier).

$$\beta=\frac{\sqrt{ab}}{\omega_c}=0.0671, \qquad \alpha=0.2.$$

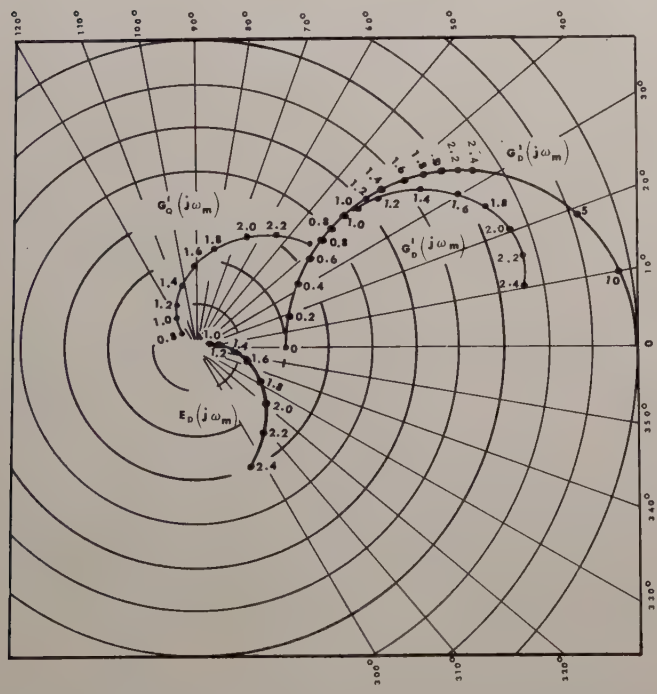


Fig. 12—Network functions for wide-band case (60-cps carrier). (Parameter is relative frequency $u=\omega_m/\sqrt{ab}$).

Specifications

Attenuation	=5:1
Lower corner (a)	=12 cps
Upper corner (b)	=60 cps
Maximum phase shift	=41.8° at 26.8 cps

$\sqrt{ab}$=168.3 rad/sec, $\beta=\sqrt{ab}/\omega_c$=0.447, α=0.2.

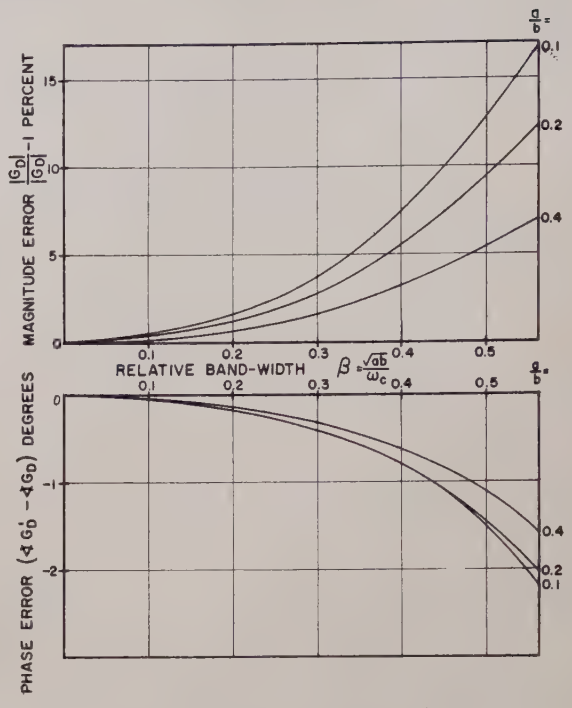


Fig. 13—Approximation error at $\omega_m=\sqrt{ab}(u=1)$, desired frequency of maximum phase shift.

Concluding Remarks

A reasonably complete theory of network behavior for carrier-suppressed modulation is now available, including a logical synthesis method. To a large extent, however, the theory and particularly the synthesis method are academic; the only practical network which can be synthesized for conventional servo compensation is the bi-quadratic lead network. The bi-quadratic lag network is impractical because of the fantastically large coil Q's required; higher order functions are impractical because several tuned circuits must be kept aligned. This whole formidable theoretical apparatus is then used to derive just one single network. To all this must be added the problem of carrier frequency drift.

On the positive side, there now exists a sound basis for the empirical methods developed in the past. The so-called "narrow-band" approximation concept is seen to be a misnomer on more than one account. First, the approximation is excellent over a wide frequency band; secondly, realizability requirements demand a network with frequency dependence stretched out over a wide band of frequencies. "Low-pass approximation" would be a better designation.

Acknowledgment

I am indebted to Dr. J. G. Truxal for his untiring helpfulness and enthusiasm. Without his support this work could not have been carried out. I have also benefited from numerous discussions with my colleagues, particularly H. Levenstein and Dr. M. Panzer. Lastly, I particularly want to thank H. Ur for his assistance with the root locus problems.

Design Aspects of Attitude Control Systems*

M. F. MARX†

Summary—Figures of merit, besides those of performance, are discussed relative to the attitude control system of a vehicle capable of leaving and returning to the atmosphere. In addition to extreme changes in flight condition, these applications are subject to variations in configuration and performance requirements.

Traditionally, control optimization has been concerned with minimizing or maximizing a variable system function such as error. Quite often these error criteria are replaced by other criteria such as invariance and the capacity for adaptability. In fact, during a complete mission including exit to re-entry it is desirable to utilize variable figures of merit.

Examination of the control requirements of a modern returnable space vehicle illustrates how the various figures of merit dictate the design configuration. In those phases of the mission where self-adaptive control is employed, the figure of merit is usually determined by the particular technique selected. It is further demonstrated how the figure of merit varies with the mission phase as the control actuation configuration changes.

I. Introduction

THE literature of recent years contains much information concerning elements needed to synthesize attitude control systems. Much progress has been made along the lines of control system techniques, particularly in the areas of optimum and self-adaptive techniques. The pursuit of knowledge in these areas has done much to refine methods of analysis of linear and nonlinear systems.

Parallel efforts have been made in the fields of actuation or "muscles" and sensors for attitude control systems.[1] Methods are available for the application of control torques ranging from several inch-ounces to many foot-pounds.

This paper indicates some of the trade-off studies which are necessary to select the actuation means, the sensors, and the computations required to achieve satisfactory attitude response. With the large store of information available on actuation and control techniques the proper choices are not simple. The fact that many modern vehicles, particularly those of the space variety, span large performance ranges results in a requirement for the combination of more than one attitude system and for methods to switch control from one system to the other.

One reason for preparing this paper is to focus attention on the many factors leading to the control configuration besides performance. The procedure adopted by a number of investigators has been to select a performance criterion or figure of merit such as a mean-square error criterion and then by analytic means to derive a system which satisfies this criterion. The system which best satisfies the selected criterion is, by definition, optimum. If this approach is adopted, it is evident that the performance criterion, in reality, determines the system configuration. But who is to say that the proper performance criterion has been selected? It is hoped that the examples contained herein will point out additional considerations which may be useful in arriving at the optimum performance criterion.

II. Mission Phases

For purposes of illustration, an arbitrary mission which is broad enough to encompass the major control types has been selected. It is assumed that the attitude control system for a recoverable glide vehicle having orbital capability must be synthesized. The mission phases of the example selected are: boost, orbit injection, orbit, orbit ejection, re-entry, and landing. Fig. 1 presents the mission phases in terms of the flight path.

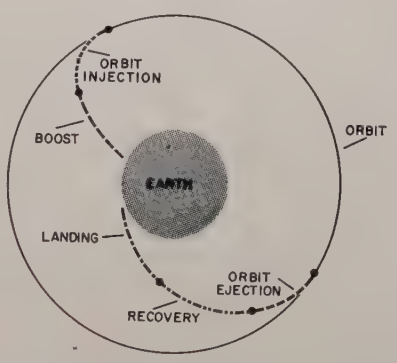


Fig. 1—Mission phases in terms of the flight path.

Using currently available booster fuel and engine combinations, and realizing that a gliding, recovery vehicle most likely will be manned, we can safely assume that at least three vehicle stages will be required to achieve orbit velocity (approximately 25,000 feet/sec for 200-mile altitude). The first stage is employed during the boost phase. The next stage and possibly also the third stage are required to achieve orbit velocity of the payload (the gliding vehicle). The orbit phase for this generalized example involves an indefinite number of orbital revolutions around the earth; a specific mission would dictate the orbital life and the attitude reference and accuracy requirements. Orbit ejection is concerned

* Received by the PGAC, May 24, 1960; revised manuscript received, October 17, 1960. Presented at the 7th Region IRE Conf., Seattle, Wash., May 25, 1960.

† Advance Space Navigation Engrg., Light Military Electronics Dept., General Electric Co., Schenectady, N. Y.

[1] W. Haeussermann, "Comparison of some actuation methods for attitude control of space vehicles," presented at Manned Space Station Symp., IAS/NASA/RAND Corp., Los Angeles, Calif.; April 1960.

with the alignment required for retrorocket firing; an energy management computer determines the firing point and the total impulse required to achieve the orbit transfer enabling satisfactory recovery. The re-entry phase of the mission deals with the altitude velocity relationships enabling satisfactory descent considering heating, acceleration, and range capability. Landing deals with the vehicle control after passing through the altitude velocity corridor; in this region, the heating danger is over and the control problem is more or less that of a conventional airplane with or without thrust capability utilizing aerodynamic control surfaces for actuation.

III. Attitude Control During Boost

During the boost phase, the vehicle configuration is similar to that presented in Fig. 2.

Neglecting speed changes, the vehicle short-period transfer function in pitch can be written

$$\frac{\dot{\theta}}{\delta} = \frac{K_\theta(1 + sT_\theta)}{\frac{s^2}{\omega^2} + \frac{2\zeta s}{\omega} + 1},$$

where

$\dot{\theta}$ = short-period pitch rate,
K_θ = short-period gain,
T_θ = path time constant,
ω = short-period resonant frequency,
ζ = short-period damping ratio,
s = Laplace differential operator.

This approximate representation of the vehicle dynamics has been selected in order not to obscure the discussion which follows. Although considerably simplified, it is a fairly good description if large angular rates and linear accelerations are avoided.

Considering the case immediately after lift off when the aerodynamic forces are negligible, the pitch rate transfer function degenerates to

$$\frac{\dot{\theta}}{\delta} = \frac{K_\theta}{s}.$$

As the velocity increases, the frequency and damping terms begin to appear in the transfer function. For the configuration shown in Fig. 2, the vehicle is statically unstable and is said to have a negative stability margin. The transfer function typically becomes

$$\frac{\dot{\theta}}{\delta} = \frac{K_\theta(1 + sT_\theta)}{\left(1 - \frac{s}{T_1}\right)\left(1 + \frac{s}{T_2}\right)}.$$

The interesting point to the control designer is that the process is dynamically unstable. Although the configuration can be stabilized by simple rate feedback, the response in general cannot be made sufficiently fast to assure adequate path control.

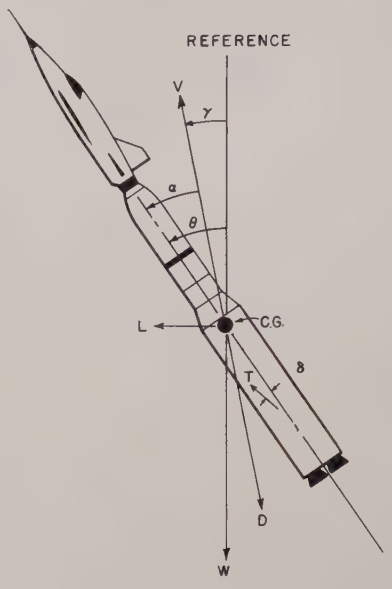


Fig. 2—Vehicle configuration during boost phase.
γ = Flight path angle measured from the reference trajectory
L = Aerodynamic lift
D = Aerodynamic drag
W = Vehicle weight
T = Engine thrust
δ = Deflection of the thrust vector
θ = Pitch angle
α = Angle of attack
C.G. = Vehicle center of gravity.

Before considering further the required feedback form, several additional elements of the open-loop transfer function will be discussed. If thrust vector control is achieved by means of a gimbaled engine, the dynamics of this actuation should be included. Usually a simple first-order lag with time constant T_A is adequate for this representation. To be realistic the effect of vehicle elasticity must be included. According to Beharrell and Friedrich[2] this adds a quadratic lead-lag term to the short-period transfer function. The combined root-locus plot for attitude rate is shown in Fig. 3. (Although it is desirable to control attitude, rate is discussed since satisfactory rate control is required to obtain satisfactory attitude response.)

The pole-zero combination shown is representative of the first bending mode of the vehicle. The higher modes can be shown to add additional pole-zero combinations which have been omitted for purposes of simplicity. Quite often the first bending mode of a large booster configuration results in structural resonance in the neighborhood of 2 to 3 cps. If control mode resonance of 0.5 cps is desired, it is evident that a challenging control problem exists.

To further complicate matters, the process is not invariant. The large amounts of fuel utilized during boost

[2] J. L. Beharrell and H. S. Friedrich, "The transfer function of a rocket-type guided missile with consideration of its structural elasticity," *J. Aeronaut. Sci.*, vol. 21, p. 454; July, 1951.

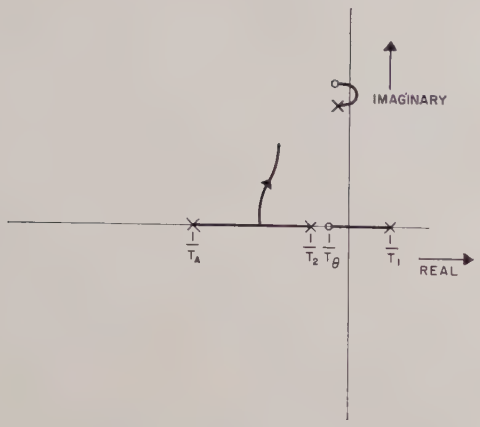


Fig. 3—Combined root-locus plot for attitude rate.

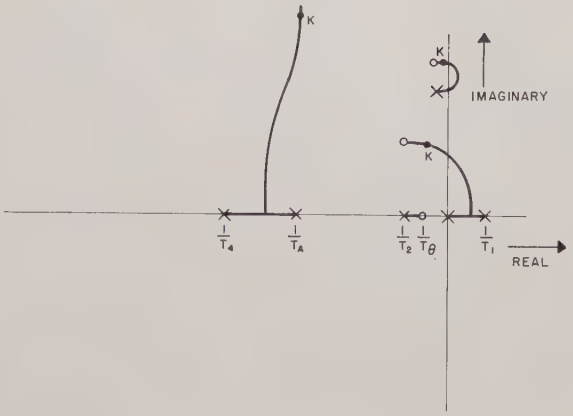


Fig. 5—Root-locus plot—feedback configuration.

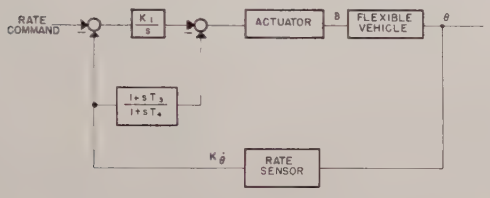


Fig. 4—High-performance rate system.

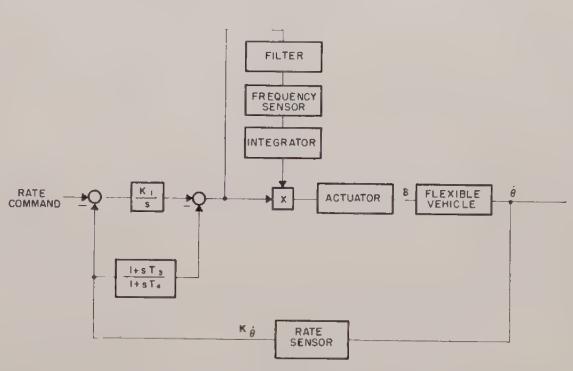


Fig. 6—Self-Adaptive rate system.

and changes in flight condition result in significant variations in the structural resonance and in the time constants T_2, T_θ, and T_1 associated with the rigid vehicle. Changes in payload usually result in similar vehicle transfer function changes.

For any fixed set of conditions the control mode must be stabilized by the addition of lead. Fig. 4 indicates a feedback configuration which accomplishes this function. An angular accelerometer or lead network is required to provide the necessary lead.

With this feedback configuration, the root locus shown in Fig. 3 changes to that shown in Fig. 5. The complex zeros in the feedback are determined by the values selected for K_1, T_3 and T_4. Error integration has been added to provide zero error steady-state response. Steady-state torque perturbations resulting from thrust misalignment and aerodynamic forces require this addition.

The resulting attitude control system is conditionally stable with respect to the structural mode. As the open loop-gain is increased from zero, the structural mode becomes unstable and again stable. If the gain is sufficiently high the structural mode will be essentially eliminated from the response and the closed-loop control mode transfer function will approach the inverse of the feedback. Thus the desired closed-loop dynamics can be obtained by the proper combination of feedback gradients provided the open-loop gain K is sufficiently high.

The system described in Fig. 4 can easily be made self-adaptive by monitoring the frequency associated with the actuator and filter time constants T_A and T_4, respectively.[3] Fig. 6 indicates the location of the gain changer shown as a multiplier.

Since the high-frequency mode is monitored, it is desired to monitor the system where this mode is predominant. A filter has been added ahead of the frequency sensor to attenuate the low-frequency components due to the control mode. If the frequency of the actuator mode deviates from the center frequency of the frequency sensor, the integrator output will cause the multiplier to change the gain so as to restore the center-frequency operation. Normally the center frequency is set as high as practical so as not to excite higher-frequency modes not shown in the representation. Thus the highest possible open-loop gain is maintained and the desired response obtained.

A continuous proportional system has been selected for control in the booster phase. The criterion selected is invariant response in preference to the error criteria discussed in the literature in connection with optimum controls and a number of the self-adaptive techniques. The primary reason for this choice is the need to tie-in with the guidance system which results in additional loops being used around the attitude control. This tie-in is greatly facilitated if the inner loops are invariant.

[3] M. F. Marx, "Recent adaptive control work at the General Electric Company," *Proc. Self-Adaptive Flight Control Systems Symp.*, January, 1959.

The continuous system choice is predicated on the presence of the lightly damped structural mode which would be needlessly excited if a discontinuous system having a limit cycle were employed.

IV. Attitude Control During Orbit

Upon achieving orbit the propulsion stages are assumed to have separated so that all that remains is the payload, which is the return component with a winged configuration. If the vehicle is manned, it is likely that the orbit will be above 100 miles altitude to be free of aerodynamic drag and under 300 miles to be clear of the Van Allen radiation belts.

For purposes of illustrating the attitude control design features, it is assumed that the satellite functions require it to preserve a fixed orientation with respect to earth vertical. This could be for reasons of communication, surveillance, or meteorology. It is further assumed that the attitude errors should be held to $\pm 0.1°$ of arc. Regarding accuracy of control, there have been specifications which have required attitude errors to be held under 0.1 second of arc. Hence it is seen that the presently assumed 0.1° specification is quite modest, relatively speaking.

Since it is required that the vehicle be slaved to the earth's vertical and the orbiting altitude lies between 100 and 300 miles, the satellite will require nearly constant angular rotation at about 4°/minute. The orbit altitude restraint requires a nearly circular orbit. Since the orbit cannot be made exactly circular, there is certain to be some orbital eccentricity. Due to this eccentricity, the satellite angular velocity will require that a sinusoidal component having a period of approximately 90 minutes be superimposed on the constant 4°/minute average rate. Hence the attitude control will be active for the entire mission phase.

In addition to the varying attitude system command caused by eccentricity of the orbit, disturbing torques due to gravity gradient, magnetic fields, solar pressure, motion of inhabitants and machinery, and aerodynamic torques from residual atmosphere require trim torques from the actuation means.

An additional important consideration entering into the design of the orbit attitude-control system is weight. Considering that it takes approximately 100 pounds of booster weight for every pound put into orbit, a trade-off study between the total control weight and the weight of the power supply must be made.

The inertial actuation means in the form of power gyros and flywheels are attractive considering the cyclic angular rate requirement resulting from orbit eccentricity. In this event the momentum interchanges between the vehicle and flywheel result in no net power requirement except that to compensate for motor and friction losses. If the flywheel must supply a trim torque it must be periodically reset by an additional torquing device such as a mass expulsion system. Hence, it is seen that the flywheel plus the required reset feature requires considerable weight but is capable of economical operation in terms of fuel requirement. This may be an important consideration for an unmanned satellite having a lifetime measured in years, but in the case of a manned vehicle where duration is relatively short, *viz.*, several days, the fuel requirement assumes a less important role.

The mass expulsion techniques such as gas reaction jets are extremely small and reliable but require expenditure of fuel mass to change the vehicle momentum. As an indication of size requirements, these units are commonly calibrated in inch-ounces. Since the resulting nozzle sizes must be extremely small if proportional control is utilized, it is advantageous to modulate with timed impulses so that relatively large jets can be used to develop the required impulse. These larger jets are less prone to plugging than the smaller jets needed for proportional control.

The timed impulse technique is well suited to the vapor-pressure type gas generator which recharges itself using solar energy. The size of the generator depends on the duty cycle to which the control is exposed and the perturbation torques encountered.

The other consideration which leads to the timed impulse method as an attractive choice in this application is that of sensor deadband which leads to unstable operation within the deadband. This is particularly a problem regarding the rate stabilization signal. At the low rates encountered with these controls it is exceedingly difficult to obtain rate sensors with thresholds low enough to sense the motion. Hence, the control technique which provides adequate stability without requiring rate is desirable.

The timed impulse technique accepts the fact that the system will limit the cycle, and uses it to full advantage. Through suitable logic the correct impulse can be applied to reverse angular direction at the switching boundary, even in the presence of a trim torque. The switching boundary is determined by the over-all accuracy requirement of the attitude control.

In addition to the difference in force levels and speed of response, another important difference between the orbit and booster control phases relative to attitude control is the difference in requirements for feedback parameter variations. While the vehicle is in orbit, the very low aerodynamic forces and the low angular rates involved make the attitude control essentially independent of velocity. Furthermore, the low fuel expenditures result in negligible changes in the vehicle total mass and mass distribution. Hence the vehicle's transfer function is virtually invariant, thus obviating the necessity of a self-adaptive system. This being the case, such factors as orbit duration, accuracy requirements, system weight, and reliability are the items which lead to the formulation of the system configuration.

V. Attitude Control During Recovery

These phases of the mission result in still another form of attitude control having characteristics somewhat similar to the booster phase but entirely different from the orbit phase. These differences are apparent if one reviews the general concept of the recovery of a lifting vehicle.

Orbit ejection is a form of orbit transfer where the transfer is from the circular or slightly elliptic orbit to a parabolic trajectory prior to reentry into the earth's atmosphere. This orbit transfer is accomplished by the proper application of a thrust impulse of sufficient magnitude and direction to impart the desired velocity increment to the vehicle. The size of the velocity increment is derived from an energy management computer which is discussed subsequently.

During the orbit ejection phase of the mission the attitude control serves to align the vehicle prior to and during the firing of the retrorocket used to impart the required velocity increment. Whether a radial or tangential firing technique is employed is immaterial to this discussion since accuracy and force-level differences are not sufficiently large to determine the attitude control technique.

The performance specifications on the attitude control depend on the allowable variations in magnitude and direction of the desired velocity increment used for ejection. This increment is dictated by the type of reentry employed by the re-entry computer. Generally speaking, attitude control requirements prior and during retrorocket firing are in the neighborhood of 0.1°. If the vehicle is slaved to the earth's vertical during orbit there is no firm requirement for large slewing rates, but the control must have adequate authority to compensate for variations in thrust direction and displacement relative to the center of gravity of the vehicle. These thrust asymmetries usually range within 0.1° in alignment and 0.1 inch in displacement depending on the engine size and design. Engines utilizing solid propellants are usually more difficult to align than those employing liquid propellants.

The trim requirements due to thrust misalignment result in an attitude control whose authority is several orders of magnitude larger than that required for orbit control where the perturbations are infinitesimal in comparison.

At this point it is reasonable to question why this powered phase should be different from the boost phase. The difference is that during ejection the thrust is applied for a relatively short time after which a ballistic trajectory is followed prior to re-entry. The rest of the flight, at least for presently contemplated designs, is unpowered due to weight restrictions. Hence, the high-level control used for orbit ejection will also be used for alignment prior to re-entry where dynamic pressure is too low for aerodynamic control. Later in the flight after dynamic pressure is sufficiently high, aerodynamic actuation is feasible.

As indicated previously, the recovery attitude-control accuracy specifications are determined by the energy management computer which for this discussion can be viewed as a guidance computer. It functions to generate the proper steering commands to the control system so that the vehicle recovery is made within the constraints of allowable heating, acceleration limits of the vehicle and inhabitants, and dispersal of the landing sites. Certain variations of the problem, such as a minimum time recovery and an alternate landing site designation, are handled by the computer.

The computer problem is mainly one of prediction due to the limited maneuverability of the vehicle and the finite amount of energy associated with unpowered flight. Although the energy management computer is another subsystem, it is mentioned at this time to point out the close relationship between the two systems. One cannot be designed without full consideration of the other.

One of the most common high-level nonaerodynamic control actuation techniques presently considered is the mass expulsion system powered by a cold gas or hot gas power supply. The control is required for only a very limited time so the hydrogen peroxide systems such as employed on the X-15 research airplane or the ammonium nitrate powder grains offer neat light packages.

The use of hot gas power supplies transfers the design problem to the systems engineer. Due to contamination and high temperature it is difficult to modulate the high-level controls to obtain proportional action and next to impossible to throttle to zero. For these reasons the discontinuous control techniques employing base-width pulse modulation or conventional relay control are suitable. As an example of the later technique, the schematic shown in Fig. 7 illustrates such a mechanization.[4]

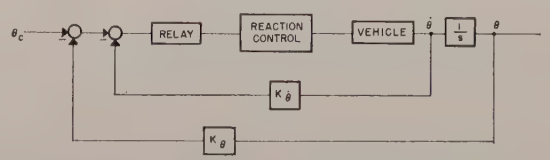


Fig. 7—Discontinuous control technique.

These systems are characterized by a limit cycle due to the dead zone of the relay or bistable element. They are sometimes referred to as hysteresis systems. The limit cycle amplitude and frequency are functions of the open-loop gain and lead obtained from the rate feedback. The hysteresis of the relay is usually a destabilizing factor. Stability analysis of these systems is com-

[4] R. L. Cosgriff, "Nonlinear Control Systems," McGraw-Hill Book Co., Inc., New York, N. Y.; 1958.

monly carried out employing phase plane techniques if the system order is not greater than two. Higher order systems necessitate the use of phase space techniques. These procedures, however, are so hard to visualize, that computing facilities are usually utilized in these instances.

The high-level controls used in recovery generally reflect the use of techniques different from the timed impulse method formed in the low-level orbit controls. The control mode frequencies are typically in the range of three to six radians per second. Consequently, rate gyros or lead networks can be successfully applied as in conventional flight control applications.

Earlier in the discussion it was mentioned that the first part of the recovery required the use of nonaerodynamic control until sufficient aerodynamic control was available so that aerodynamic control surfaces regulated by a self-adaptive control system could be used. It is desired to make the changeover to aerodynamic control as early as possible in order to minimize the amount of reaction fuel required. The change from one control type to the other, and their combined use, is called control blending. The same problem is encountered in control system design for vertical landing or takeoff aircraft (VTOL) during transition from hovering to conventional flight.

Although blending, when first considered, appeared difficult to accomplish automatically, the solution was rather simple. Fig. 8 illustrates a combined aerodynamic and nonaerodynamic system.

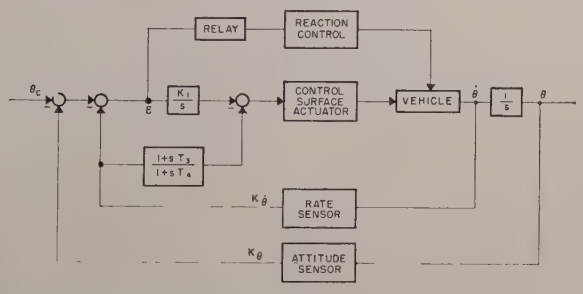


Fig. 8—Combined aerodynamic and nonaerodynamic attitude control.

The aerodynamic control portion of the system is similar to that described for booster control. The system error ϵ controls the firing of the reaction control. Whenever the aerodynamic control loses effectiveness as indicated by error buildup, the nonaerodynamic control activates automatically. It is interesting to note the operation of the combined system is such that as the least amount of aerodynamic force appears, the effective spring constant of the vehicle drives the steady-state error signal to zero, thus eliminating the limit cycle of the nonaerodynamic system. This operation is highly desirable from the viewpoint of control life and fuel economy.

Considering the attitude control requirements of the recovery phase, the performance is satisfactory if it adequately activates the steering commands generated by the guidance system, in this case the energy management computer. Typical requirements of the attitude system are zero error characteristics, which imply error integration; an equivalent resonant frequency of approximately five radians per second; and a damping ratio of 0.7.

As in the case of the booster control, the attitude control and guidance system compatibility are important considerations. The integration of these subsystems is greatly facilitated if the attitude transient response is invariant, rather than if it satisfies an arbitrary error criterion.

The absence or presence of atmosphere dictates the form of actuation, aerodynamic or nonaerodynamic. The form of actuation in turn places constraints on the type of system which can be utilized.

Due to the high rates of response (high compared to the orbit phase), there is considerable choice available for the generation of an attitude rate signal for stabilization. Hence, in this case, the sensor requirements are not strong factors in the design of the attitude system. The design of the energy management computer, however, depends strongly on sensor availability.

Human factors are important aspects of the attitude control system design problem. Since the vehicle will be under the management of a pilot, the pilot and vehicle must be compatible. Generally this places a further specification on the system band-pass. Since the pilot's desired response can be specified within rather narrow limits, invariant attitude system response is desirable. A further requirement resulting from the presence of the pilot is that of display provision. This aspect, however, is a science in itself and is not discussed herein.

Since during the recovery phase and particularly during re-entry, the vehicle traverses extremely large ranges of altitude and velocity in short time spans, it is necessary that the control system have the capability for making very rapid changes in gain. Although this problem is more severe for a ballistic or nonlifting reentry, the gliding vehicle assumed herein also experiences difficulty. The assumed vehicle attains maximum dynamic pressure about thirty minutes after firing the retrorocket. Consequently, an adaptive or nontailoring system is most appropriate for recovery.

VI. Conclusion

The attitude control requirements of a manned lifting vehicle having orbit capability have been examined. Based on the performance requirements, integration factors, characteristics of the sensors and actuators, duration of the mission, perturbations, structural rigidity, vehicle size, and reliability, a typical control system has been described for the various phases.

For the particular mission selected, four distinct types of attitude control are advisable. During the boost phase a continuous information, self-adaptive system employing a rotatable rocket engine thrust means is

satisfactory. The requirements of the particular orbit mission selected can be fulfilled with a mass ejection system of low-authority control by timed impulses. Recovery requires the blending of nonaerodynamic control of high authority with conventional aerodynamic control. Thus the hysteresis type ON-OFF control and the continuous self-adaptive control, respectively, are appropriate.

From a philosophical point of view, it is interesting to note that the peculiarities of the attitude control problem have led to a required control configuration which is not optimum in the sense of satisfying any particular error criterion. The optimum configuration is generated as a result of a trade-off study considering many factors of which performance is only one.

VII. Acknowledgment

The author acknowledges the assistance of his associates from whose work areas he has drawn heavily; in particular, the ideas on nonaerodynamic control contributed by J. D. Welch and J. M. Cooper of the General Electric Company are appreciated.

Analysis and Design of Feedback Systems with Gain and Time Constant Variations*

KAN CHEN†, ASSOCIATE MEMBER, IRE

Summary—**The design of a feedback control system containing an element with proportional variation of gain and time constant is a common problem encountered by control engineers in practice. The problem includes the stabilization of a system, which is open-loop unstable when both the gain and the time constant of the element are negative. This paper presents a method for analyzing the transient response of systems containing the aforementioned element, and designing the systems to meet transient specifications.**

Introduction

THIS paper is concerned with the analysis and design of feedback control systems (shown in Fig. 1), which contain an element subject to proportional variation of gain and time constant. Thus, the element has a transfer function with the following form,

$$\frac{\frac{K}{a}}{\frac{T}{a}s+1}, \tag{1}$$

where K and T are the nominal values of the gain and the time constant, respectively. The parameter a may vary with load or environmental conditions at a rate much lower than the system transients. The design of such a system is usually more difficult than that of a system containing fixed parameters. This is particularly so when a becomes negative ($a<0$), in which case the system is open-loop unstable.

* Received by the PGAC, February 2, 1960; revised manuscript received, November 10, 1960. This paper was presented at the 1960 WESCON Convention at Los Angeles, Calif.

† Westinghouse Electric Corp., Pittsburgh, Pa.

Although the transfer function (1) appears to represent only a limited class of elements, such elements are frequently encountered in practice. A few examples will be enumerated below. It will also become obvious that the method to be presented can easily be extended to study the effects of slow variations of several parameters in a system.

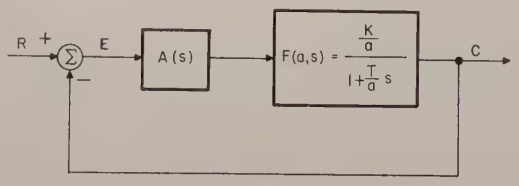


Fig. 1—Feedback loop including element subject to gain and time-constant variations.

Practical Examples

A. Rotating Generator Under a Wide Temperature Range ($a>0$)

In controlling the output voltage of a generator, the field current is varied by changing the applied voltage to the field. The voltage gain of the generator is therefore inversely proportional to the field resistance. The same is true for the time constant of the field winding, since it is equal to the ratio of its inductance to its resistance. Consequently, the voltage gain and the time constant will vary in direct proportion as the field resistance varies with temperature. The frequency response function of the generator is hence described by (1), in which $a>0$ is proportional to the slowly varying field resistance.

B. Open-Loop Instability in Reactor Controlled AC Motor Drives

In a reactor-controlled ac motor-drive system [1], the developed motor torque τ is dependent upon the reactor control current I_c. The speed (n)-torque (τ) characteristics are shown as a family of curves in Fig. 2, each curve corresponding to a certain value of reactor control current. For a small range of operation, the dynamic characteristics of the linearized system are represented by the block diagram in Fig. 3. The transfer function between motor speed and reactor control current is in the form of (1), with

$$K = \frac{\partial \tau}{\partial I_c},$$

$$T = J \text{ (inertia)},$$

and

$$a = -\frac{\partial \tau}{\partial n}.$$

Obviously, the value of a is negative at the operating point P_2, shown in Fig. 2. In other words, the open-loop system (I_c=constant) is unstable, since the slightest increase (or decrease) in speed would cause additional acceleration (or deceleration). This is always the case when the system tries to regulate at low speed.

Other examples which can be cited are control systems using magnetic amplifiers within wide temperature ranges, voltage control of self-excited alternators under capacitive loads, and speed control of steam turbines within wide speed ranges.

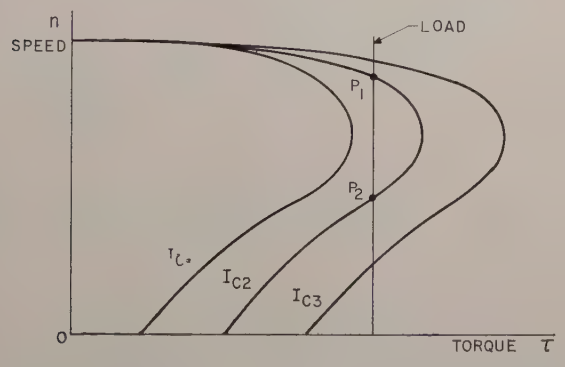


Fig. 2—Speed-torque characteristics of ac motor drives.

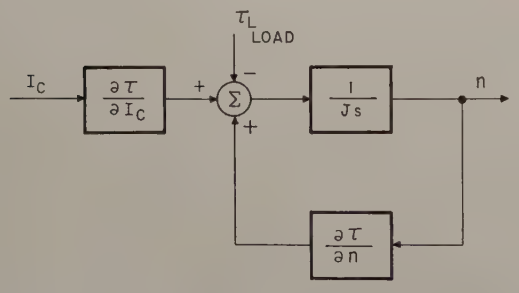


Fig. 3—Linearized block diagram of ac motor-drive speed-control systems.

Relationship Between Open-Loop Bode Plot and Closed-Loop Transient Response

It has been found that the gain asymptotes of the open-loop Bode plot can be used to determine quickly the approximate locations of the closed-loop poles [2]. The rules for the approximate evaluation and the effect of the closed-loop poles on transient response are as follows:

1) The open-loop poles which correspond to asymptote breaks below the −15-db line are also approximate closed-loop poles. These poles give small and rapidly decaying transient terms, which have negligible effect upon the main transient response.
2) The open-loop zeros which correspond to asymptote breaks above the +15-db line are the approximate closed-loop poles. These poles give small but slowly decaying transient terms, which have insignificant effect upon the main transient response, but may add a long tail to the system transient.
3) The asymptote breaks inside the ±15-db band determine the approximate location of the dominant closed-loop poles, which in turn determine the main system transient.

These rules constitute the basis of the methods presented in this paper.

Method of Analysis

Consider the feedback control system shown in Fig. 1. It is assumed that the only varying parameter in this system is a.

A. Positive a

If the variable parameter a remains positive, the above rules may be used without modification. The Bode gain asymptotes for the transfer function in (1) are shown in Fig. 4. Notice that as a varies, the asymptote with −20 db/decade slope is fixed in position. Only the horizontal asymptote moves with varying a. If a two-time constant controller is considered,

$$A(s) = \frac{A_0}{(T_1 s + 1)(T_2 s + 1)}. \tag{2}$$

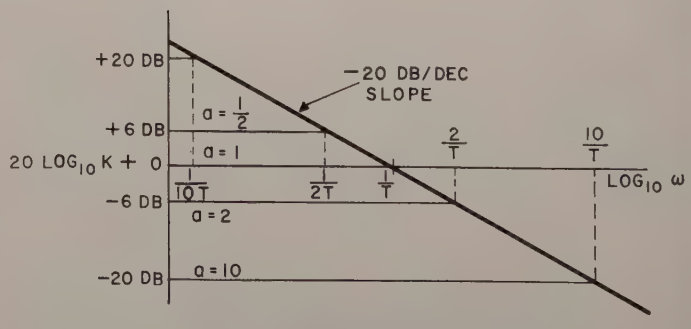


Fig. 4—Bode plot of the element with proportional variation of gain and time constant.

In this case, the Bode plot is like that shown in Fig. 5. If the gain (A_0K/a) remains greater than +15-db and if the time constant (T/a) remains greater than T_1 and T_2, the Bode plot inside the ±15-db band will not be affected and, therefore, the main system transient, as well as the system stability, will be relatively unaffected by the variation of a. This is an important result. It indicates a way to make the system insensitive to the varying parameter a.

If the time constant T/a is relatively small and its corresponding break on the Bode plot is below the −15-db line, as shown in Fig. 6, the variation of a will affect the Bode plot within the ±15-db band. However, since the only break dependent on a is below the −15-db line and therefore corresponds to a negligible transient term, the parameter a affects the system performance just as much as the loop gain (A_0K/a) does. In other words, the main system transient depends upon a mainly through its effect upon the gain alone. Thus, it is usually true that the smaller the value of a, the faster the system responds and the more the system tends to be unstable. As has been shown in a previous paper [2], the dominant closed-loop poles can be approximately determined by considering the open-loop transfer function corresponding to the Bode plot inside the ±15-db band. Referring to Fig. 6, this means that the open-loop pole at $-1/T_1$ is substituted by an open-loop pole at the origin, and the open-loop pole at $-a/T$ is simply ignored. In other words,

$$G(s) \approx \frac{\frac{A_0K}{T_1a}}{s(T_2s+1)}. \tag{3}$$

Therefore, the *approximate* dominant closed-loop poles will vary with a as shown by the root locus in Fig. 7.

When the asymptote break dependent on a is within the ±15-db band, the main system transient depends upn a through its effect on both the gain and the time constant T/a. When this is the case, the method of "root contours" [3] may be used to determine the main system transient. For example, if the Bode plot is like that shown in Fig. 8, the open-loop transfer function is approximately [2]

$$G(s) \approx \frac{\frac{\omega_c}{a}}{s\left(\frac{T}{a}s+1\right)}, \tag{4}$$

where ω_c is the crossover frequency, the frequency at which the Bode asymptotes cross the 0-db line. The expression in (4) is obtained by ignoring all the breaks outside the ±15-db band. Thus, the corresponding approximate closed-loop transfer function is

$$\frac{G}{1+G} \approx \frac{\omega_c}{Ts^2+as+\omega_c}, \tag{5}$$

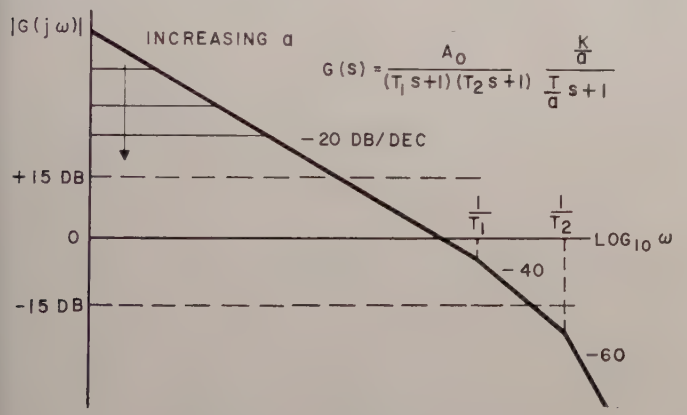


Fig. 5—Bode plot for large T/a.

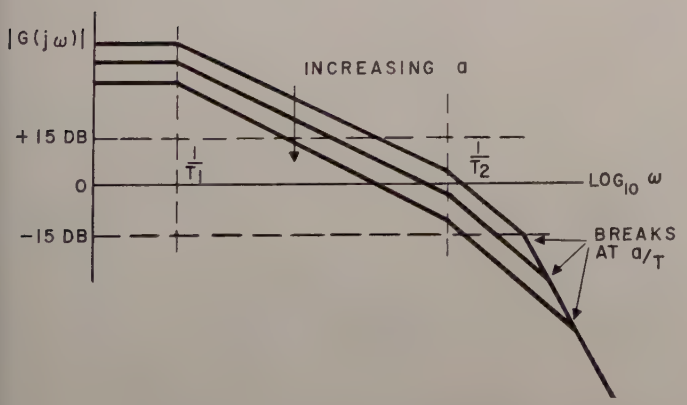


Fig. 6—Bode plot for small T/a.

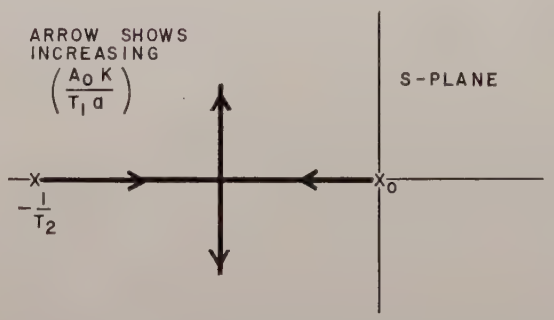


Fig. 7—Root locus of dominant closed-loop poles for the case of small T/a.

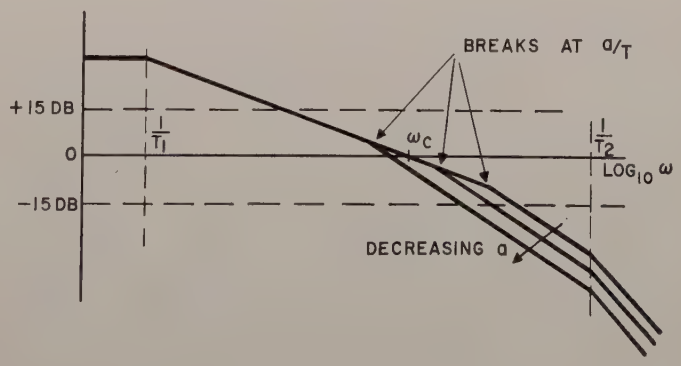


Fig. 8—Bode plot for medium T/a.

and the corresponding approximate dominant closed-loop poles are

$$s_p \approx \frac{-a \pm \sqrt{a^2 - 4T\omega_c}}{2T}. \tag{6}$$

The root contours (loci of s_p for various values of a) are shown in Fig. 9. Again, decreasing a tends to make the system respond faster but become more unstable.

Although the examples given above are for a second-order controller, it is clear that the method of analysis is good for systems of any order.

B. Negative a

When the value of a is negative, the open-loop system is unstable, and the rules given previously for the approximate evaluation of closed-loop poles from the Bode plot are no longer applicable. However, the same rules may be used if the following approximation is made:

$$\frac{\frac{K}{a}}{\frac{T}{a}s+1} = \frac{K}{Ts+a}$$

$$= \left[1 - \left(\frac{a}{Ts}\right) + \left(\frac{a}{Ts}\right)^2 - \left(\frac{a}{Ts}\right)^3 + \cdots\right]\frac{K}{Ts}$$

$$\approx -\frac{Ka}{(Ts)^2}\left(-\frac{T}{a}s+1\right) \quad \text{for} \quad \left|\frac{a}{Ts}\right| \ll 1. \tag{7}$$

Now, for *negative* values of a, every term on the right-hand side of (7) is positive and the rules given previously for the Bode plot analysis will be valid. The approximation is good, provided that the frequency under consideration is high as compared to $|a|/T$, so that the higher order terms of a/Ts are negligible.

Another way to see the validity of approximation (7) is to compare the frequency response characteristics of the exact expression with those of the approximate expression. With s replaced by $j\omega$, the exact expression on the left-hand side of (7) has the following amplitude and phase angle:

$$\text{Amplitude} = \frac{K}{\sqrt{a^2 + (T\omega)^2}} = \frac{K}{a}\frac{1}{\sqrt{\frac{(T\omega)^2}{a} + 1}}; \tag{8}$$

$$\text{Phase} = -\arctan\frac{T\omega}{a}. \tag{9}$$

The approximate expression on the right-hand side of (7) has the following amplitude and phase angle:

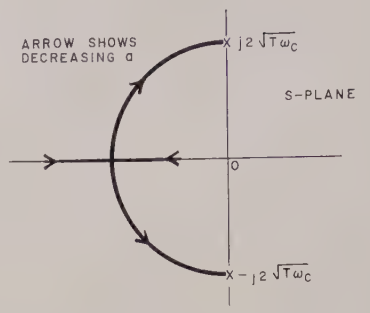


Fig. 9—Root contours of dominant closed-loop poles for the case of medium T/a.

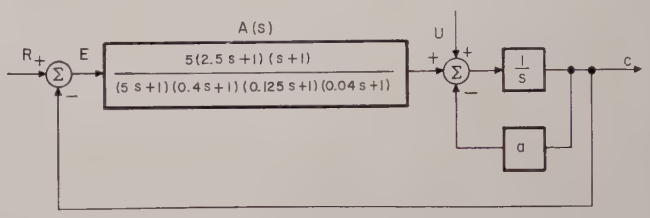


Fig. 10—A fifth order system containing an element with proportional variation of gain and time constant.

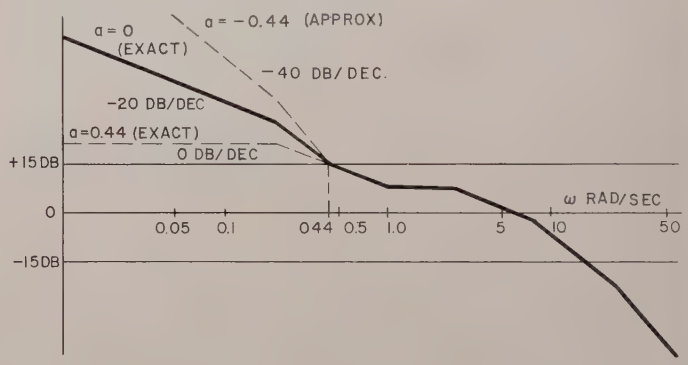


Fig. 11—Bode plot analysis of the fifth order system in Fig. 10.

$$\text{Amplitude} = \frac{Ka}{(T\omega)^2}\sqrt{\frac{(T\omega)^2}{a} + 1}; \tag{10}$$

$$\text{Phase} = -\arctan\frac{T\omega}{a}. \tag{11}$$

It is seen that the exact and approximate expressions have identical phase angles at all frequencies. Although their amplitudes are different at low frequencies, when $T\omega/a \gg 1$, both amplitudes approach $(K/T\omega)$ and are therefore approximately equal to each other.

It is recalled that the main system transient is closely correlated to the Bode plot within the ±15-db band or near the crossover frequency ω_c. Thus, if the crossover frequency ω_c of the system is much higher than $|a|/T$, or the break due to $|a|/T$ is above the +15-db line, then the approximation (7) is good for estimating the major system transient performance. Furthermore, if $|a|/T$ is kept small enough so that its corresponding break on the Bode plot (using the exact expression for

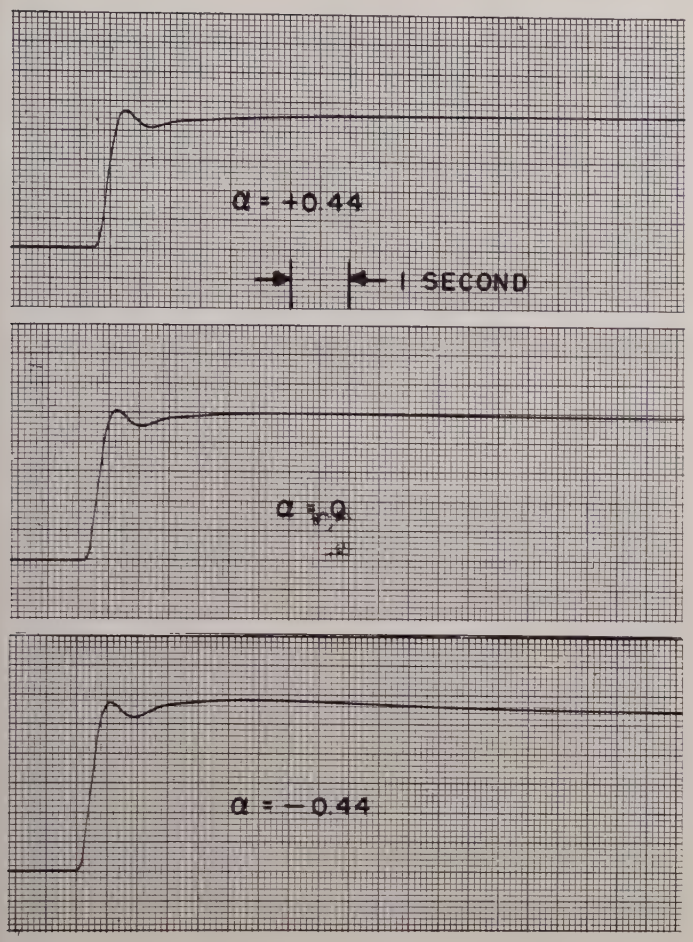


Fig. 12—Exact transient responses of the fifth-order system in Fig. 10

positive a and the approximate expression for negative a) is always above the +15-db line, then the major system transient will not be affected very much by the variation of a. The following example illustrates this point.

Fig. 10 shows a fifth-order system with the parameter a to vary from positive to negative values. In this case, T is equal to unity. The Bode plot of the system with $a=0$ is shown by solid lines in Fig. 11. It is seen that the Bode plot crosses the +15-db line at $\omega=0.44$ rad/second. Therefore, if a is less than 0.44, the Bode plot within the ±15-db band will not be affected by a (see the dotted lines in Fig. 11), and the major system transient response is essentially the same. This is evidenced by the system responses to a step input shown in Fig. 12. The three transients in this figure were obtained on an analog computer and correspond to $a=+0.44$, 0 and -0.44, respectively. Notice that the overshoot and rise time in all three cases are almost identical. Fig. 13 shows the analog computer results of the system response for $a=-0.44$, when the portion of the system involving a is replaced by the approximation given by (7). The comparison of this transient with the third one in Fig. 12 shows how good the approximation is.

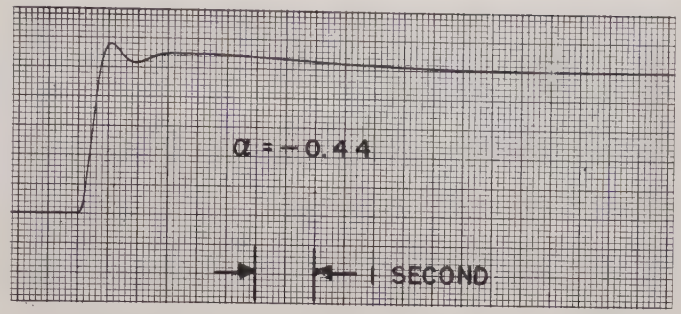


Fig. 13—Approximate transient response of the fifth-order system in Fig. 10.

Method of Design

The analysis given in the last section sheds light upon the principle of design for feedback control systems containing varying gain and time constants. In order that the system transient response will be essentially unaffected by the variation of the parameter a, the designer should exercise his freedom by choosing the controller transfer function $A(s)$ such that the break due to T/a in the Bode plot lies above the +15-db line.

Example 1. (Refer to Fig. 1).

Given:

$$F(s)=\frac{\dfrac{1}{a}}{\dfrac{0.5}{a}s+1},$$

where

$$-2\leq a\leq +4.$$

Specified:

$|E/R|$ (at zero frequency) ≤ 0.1.
Bandwidth: between 50 and 150 radians/second.
Damping ratio of dominant closed-loop poles ≥ 0.5.

To find: $A(s)$.

The three specifications given above are related, respectively, to the desired accuracy, response time, and relative stability (or overshoot of response to step input) of the closed-loop system. Since the open-loop Bode plot within the ±15-db band determines the bandwidth and the damping ratio of the dominant poles, the designer should shape this part of the open-loop Bode plot to meet the closed-loop system specifications. Thus, in Fig. 14, we obtain the asymptotes (b) and (c) within the ±15-db band. Notice that the bandwidth of the closed-loop system will be approximately equal to the crossover frequency (frequency at which the Bode asymptotes cross the 0-db line) at 100 radians/second. The break at 100 radians/second is chosen so that the transfer func-

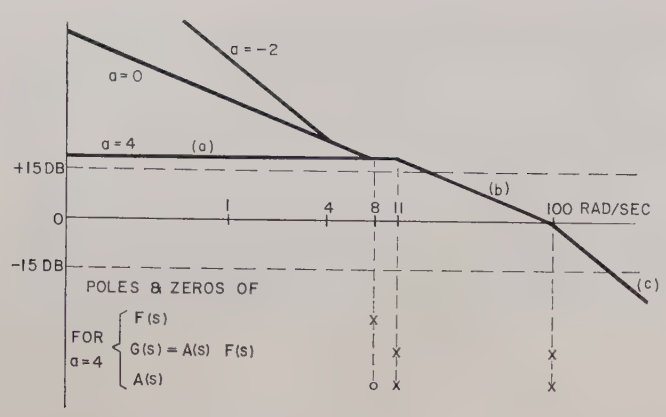


Fig. 14—Design on Bode plot for example 1.

tion inside the ±15-db band is approximately given by

$$G(s) \approx \frac{100}{s\left(\frac{s}{100}+1\right)}, \tag{12}$$

and

$$\frac{G}{1+G} \approx \frac{100}{\frac{s^2}{100}+s+100}. \tag{13}$$

Thus, the damping ratio of the approximate dominant closed-loop poles is 0.5.

If the gain of the controller is A_0, the ratio of error E to reference R is given by

$$\frac{E}{R} = \frac{1}{1+\frac{A_0}{a}} = \frac{a}{a+A_0}. \tag{14}$$

The worst case corresponds to $a=4$. To meet the accuracy specification of $|E/R| \leq 0.1$,

$$A_0 = \frac{(1-0.1)4}{0.1} = 36, \tag{15}$$

or a loop gain of 9. This fixes the low-frequency asymptote (a) in Fig. 14 for the case of $a=4$. Using the asymptotes (a), (b) and (c) in Fig. 14, we have the following transfer function for the controller:

$$A(s) = \frac{36\left(\frac{s}{8}+1\right)}{\left(\frac{s}{11}+1\right)\left(\frac{s}{100}+1\right)}. \tag{16}$$

The zero at $s=-8$ is used to cancel the pole of $F(s)$ at the same location for the case of $a=4$.

Now the effect of varying a is studied. Using the method given in the last section, the asymptotes corresponding to $a=0$ and $a=-2$ are obtained as indicated in Fig. 14. It is noticed that the Bode plot within the ±15-db band is unaffected by a, and therefore the given specifications are met regardless of the variation of a.

Example 1 is a relatively easy design problem since, fortunately, the break due to T/a can easily be restricted to above the +15-db line. It is clear that this would not be the situation if the specified bandwidth for the closed-loop system should be considerably narrower, for reasons of noise rejection and less requirement of maximum available forcing in the power amplifiers. In this case, the Bode plot within the ±15-db band will be affected by a. A minor feedback loop will then be desirable to limit the variation of the Bode plot within the ±15-db band.

Example 2

The design problem is the same as in Example 1 except that the bandwidth is expected to be between 3 and 9 radians/second.

A quick try of the approach used in Example 1 shows that the same approach will not work in this example, since the break frequency due to T/a (equal to 8 radians/second for $a=+4$) is in the order of the specified bandwidth. Thus, we try to use the Bode plot configuration shown in Fig. 6. In this figure, it is noticed that the bandwidth of the closed-loop system changes in direct proportion to a. Besides, the approximation used in (7) cannot be applied for the frequency range near the bandwidth. One way to get around this problem is to use minor loop feedback around the given controlled element, as in Fig. 15. Now the minor closed-loop trans-

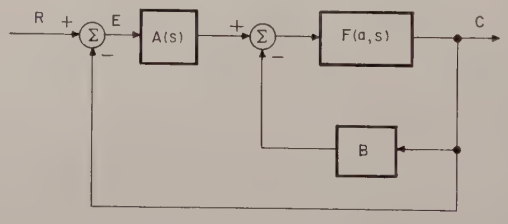


Fig. 15—Minor loop feedback for example 2.

fer function is

$$\frac{\frac{1}{a+B}}{\frac{0.5}{a+B}s+1} = \frac{\frac{1}{a_e}}{\frac{0.5}{a_e}s+1}, \tag{17}$$

where $a_e=a+B$ is the equivalent a of the minor loop. Recall in the analysis of Fig. 6 that as a increases, the bandwidth will decrease proportionally. Thus, to limit the bandwidth variation to within 3 to 9 radians/second as specified, we shall choose B such that a_e varies by only a factor of 3, for $-2 \leq a \leq 4$. That is,

$$\frac{4+B}{-2+B} = 3. \tag{18}$$

Hence,

$$B = 5.$$

Now we can design the controller transfer function $A(s)$. We shall first consider the case of $a = -2$ or $a_e = 3$. Since this is expected to correspond to minimum relative stability and maximum bandwidth, we obtain the asymptotes (b) and (c) in Fig. 16. These asymptotes

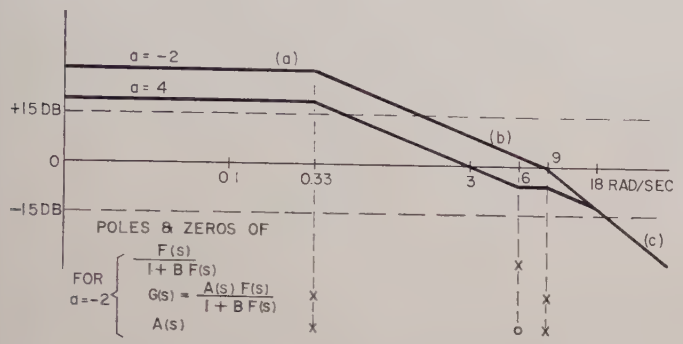


Fig. 16—Design on Bode plot for example 2.

give a bandwidth of 9 radians/second and damping ratio of dominant closed-loop poles equal to 0.5. Anticipating that the loop gain will reduce by a factor of 3 when a_e takes the opposite extreme value, the low-frequency asymptote (a) in Fig. 16 is chosen to give a loop gain of 27. Using the asymptotes (a), (b), and (c) in Fig. 16, we obtain the following controller transfer function:

$$A(s) = \frac{81\left(\frac{s}{6}+1\right)}{\left(\frac{s}{0.33}+1\right)\left(\frac{s}{9}+1\right)}. \qquad (19)$$

As shown in Fig. 16, when $a_e = 9$ (or $a = 4$), the bandwidth will decrease to 3 radians/second, the loop gain to 9, and the system will be relatively more stable (higher damping ratio). Therefore, the system meets all specifications, regardless of the variation of a.

Conclusion

A method of analyzing and designing feedback control systems containing varying gain and time constant has been presented. The method is based mainly on a relationship which has previously been established between transient response and Bode plot. The method has been shown to be applicable also to open-loop unstable systems.

Although the variations of gain and time constant are assumed to be directly proportional in this paper, as is found in many practical cases, the author feels that the general approach given in this paper can be readily extended to the analysis and design of systems containing other types of parameter variations.

Acknowledgment

The author is indebted to Dr. W. Leonhard for several valuable discussions pertaining to the development of the materials in this paper.

References

[1] W. Leonhard, "Elements of reactor controlled reversible induction Motor Drives," *Trans. AIEE*, vol. 78, pt. 2, pp. 106–115; March, 1959.
[2] K. Chen, "A quick method for estimating closed-loop poles of control systems," *Trans. AIEE*, vol. 76, pt. 2, pp. 80–87; 1957.
[3] V. C. M. Yeh, "Synthesis of feedback control systems by gain-contour and root contour," *Trans. AIEE*, vol. 75, pt. 2, pp. 85–96; 1956.
[4] G. A. Biernson, "Quick methods for evaluating the closed-loop poles of feedback control systems," *Trans. AIEE*, vol. 72, pt. 2, pp. 53–70; 1953.

Evaluation of Transient Response Coefficients*

D. S. BILLINGSLEY† AND M. G. REKOFF, JR.‡, MEMBER, IRE

Summary—**A method is presented for evaluating the transient response coefficients of response functions having poles of any multiplicity. The calculations are effected by use of a recursion-type formula with data obtained from distance and angle measurements on the root-locus plot of the system being analyzed.**

Introduction

ONE of the great advantages in the use of the root-locus method is the ability to determine graphically the coefficients for the transient response of the closed-loop system to various forms of excitations such as unit impulse, unit step and the unit doublet. The procedure in the absence of multiple roots is straightforward and is described in several [1]–[3] publications. If, however, one wishes to investigate systems with higher-order excitations as unit ramp or unit acceleration, or, if one wishes to examine functions having roots of multiplicity greater than one, then the conventional techniques must be extended. Murphy [1] has illustrated a procedure for obtaining the transient response coefficients via graphical measurements for response functions having poles of multiplicity two. It is the authors' purpose to extend Murphy's procedure so that response functions of any multiplicity m can be determined thereby. The method to be presented does not remove the inherent difficulty of the graphical evaluation, but presents relations and outlines a logical sequence of steps to minimize the labor required. Since large multiplicities are not expected in practical analysis, the authors have included a table for determining coefficients for multiplicities of 2 through 6. It is possible to develop a recursion-type expression from which one may determine coefficients of any multiplicity m; however, the expression is too cumbersome to be of practical value, and for multiplicities of 5 or larger, it is more expeditious to continue the differentiation described below.

Coefficients Obtained by Partial Fraction Expansion

Let the expression

$$c(s) = \frac{k \prod_i (s - z_i)}{\prod_j (s - \rho_j)}$$

represent the output of the system under investigation when excited with an appropriate input. Suppose there exist one or more poles of multiplicity greater than 1 and consider one such pole ρ_w of multiplicity m; then this expression can be rewritten as

$$c(s) = \frac{k \prod_i (s - z_i)}{(s - \rho_w)^m \prod_{j \neq w} (s - \rho_j)}.$$

Expansion in partial fractions produces

$$c(s) = \sum_{n=1}^{m} \frac{\lambda_n}{(s - \rho_w)^n} + [\text{remaining terms in partial fraction expansion}],$$

where the λ_n are the coefficients associated with the pole of multiplicity m and are determined by proceeding with the partial fraction expansion obtaining

$$(s - \rho_w)^m c(s) = \sum_{n=1}^{m} (s - \rho_n)^{m-n} \lambda_n + (s - \rho_w)^m [\text{remaining terms}].$$

Using the conventional method of evaluating the λ_n, one proceeds as follows:

$$\lambda_m = \lim_{s \to \rho_w} [(s - \rho_w)^m c(s)] = \frac{k \prod_i (\rho_w - z_i)}{\prod_{w \neq j} (\rho_w - \rho_j)}.$$

To evaluate λ_{m-1}

$$\frac{d}{ds}[(s - \rho_w)^m c(s)] = \sum_{n=1}^{m-1} (m - n)(s - \rho_w)^{m-n-1} \lambda_n + \frac{d}{ds}[(s - \rho_w)^m (\text{remaining terms})].$$

Thus,

$$\lambda_{m-1} = \lim_{s \to \rho_w} \left[\frac{d}{ds} \{ (s - \rho_w)^m c(s) \} \right].$$

To evaluate λ_{m-2},

$$\frac{d^2}{ds^2}[(s - \rho_w)^m c(s)] = \sum_{n=1}^{m-2} (m - n)(m - n - 1)(s - \rho_w)^{m-n-2} \lambda_n + \frac{d^2}{ds^2}[(s - \rho_w)^m (\text{remaining terms})],$$

* Received by the PGAC, May 2, 1960; revised manuscript received, October 5, 1960.

† Dept. of Chemical Engrg., Agricultural and Mechanical College of Texas, College Station, Texas.

‡ Dept. of Elec. Engrg., Agricultural and Mechanical College of Texas, College Station, Texas.

and

$$\lambda_{m-2} = \lim_{s \to \rho_w} \left[\frac{1}{2!} \frac{d^2}{ds^2} \{ (s - \rho_w)^m c(s) \} \right],$$

from which one can obtain the general expression

$$\lambda_{m-q} = \lim_{s \to \rho_w} \left[\frac{1}{q!} \frac{d^q}{ds^q} \{ (s - \rho_w)^m c(s) \} \right] \quad 1 \leqq q \leqq m - 1.$$

Observe that in evaluating each of the λ_{m-q} one must take the qth-order derivative of $(s-p_w)^m c(s)$ with respect to s. Since

$$(s - \rho_w)^m c(s) = k \frac{\prod_i (s - z_i)}{\prod_{j \neq w} (s - \rho_j)},$$

it appears that some closed form for the derivative of ratios of products would be useful.

Expression for the Derivative of Ratio of Products

It has been shown [4] that

$$\left(\frac{\prod u_r}{\prod v_t} \right)' = \left(\frac{\prod u_r}{\prod v_t} \right) \left(\sum \frac{u_r'}{u_r} - \sum \frac{v_t'}{v_t} \right).$$

Applying this expression to the problem at hand where s is the independent variable, let

$$u_r = (s - z_r) \quad \text{and} \quad u_t = (s - \rho_t),$$

from which it follows that

$$\frac{du_r}{ds} = 1 \quad \text{and} \quad \frac{du_t}{ds} = 1.$$

One obtains

$$\left[k \frac{\prod (s - z_r)}{\prod (s - \rho_t)} \right]' = \left[k \frac{\prod (s - z_r)}{\prod (s - \rho_t)} \right] \cdot \left[\sum \frac{1}{(s - z_r)} - \sum \frac{1}{(s - \rho_t)} \right].$$

For convenience in discussion, let

$$P = \left[\frac{\prod u_r}{\prod v_t} \right] \quad \text{and} \quad Q_l = [\sum u_r^{-l} - \sum v_t^{-l}].$$

The above expression then becomes

$$P' = PQ_1.$$

To perform successive differentiations, an expression for Q_l' is required and it can easily be verified that

$$Q_l' = - lQ_{l+1}.$$

Using the above relations, the successive derivatives can be determined as follows:

$$(P)^{\mathrm{I}} = PQ_1$$

$$(P)^{\mathrm{II}} = P(Q_1^2 - Q_2)$$

$$(P)^{\mathrm{III}} = P(2Q_3 - 3Q_1Q_2 + Q_1^3)$$

$$(P)^{\mathrm{IV}} = P(Q_1^4 - 6Q_4 - 8Q_3Q_1 + 3Q_2^2 - 6Q_1^2Q_2)$$

$$(P)^{\mathrm{V}} = P(24Q_5 - 30Q_1Q_4 - 20Q_2Q_3 + 20Q_1^2Q_3 + 15Q_1Q_2^2 - 10Q_1^3Q_2 + Q_1^5)$$

$$(P)^{\mathrm{VI}} = P(Q_1^6 - 120Q_6 + 144Q_1Q_5 + 90Q_2Q_4 + 40Q_3^2 - 90Q_1^2Q_4 - 120Q_1Q_2Q_3 - 15Q_2^3 + 40Q_1^3Q_3 + 45Q_1^2Q_2^2 - 15Q_1^4Q_2).$$

Evaluation of Coefficients

Assuming that complex poles and zeros occur in conjugate pairs, one obtains

$$\lambda_m = \lim_{s \to \rho_w} [(s - \rho_w)^m c(s)] = k \frac{\prod_i (\rho_w - z_i)}{\prod_{w \neq j} (\rho_w - \rho_j)},$$

where, in general, λ_m is a complex number whose magnitude and angle can be determined directly from measurements on the root-locus diagram. The magnitude

$$|\lambda_m| = k \frac{\prod_i |\rho_w - z_i|}{\prod_{j \neq w} |\rho_w - \rho_j|}$$

is the product and quotient of "distances" between ρ_w, the root of multiplicity m, and the remaining poles and zeros of the loop transmission. The

$$\arg(\lambda_m) = \sum_i \arg(\rho_w - z_i) - \sum_{j \neq w} \arg(\rho_w - \rho_j).$$

If ρ_w is real or zero, the argument of λ_m is zero and λ_m is a real number, resulting in the time function

$$\frac{\lambda_m}{m!} t^{m-1} \epsilon^{-\sigma_w t}.$$

If ρ_w is a complex number, then the argument of λ_m is not zero and λ_m is complex. Since the complex poles and zeros occur in conjugate pairs, the complex pairs can be combined and the time function

$$2|\lambda_m| \frac{1}{m!} t^{m-1} \epsilon^{-\sigma_w t} \cos [w_w t + \arg(\lambda_m)]$$

obtained, where

$$\rho_w = - \sigma_w + j w_w.$$

Murphy [1] demonstrated the procedure to obtain

$$\lambda_{m-1} = k \frac{\prod_i (\rho_w - z_i)}{\prod_{j \neq w} (\rho_w - \rho_j)} \left[\sum_i \frac{1}{(\rho_w - z_i)} - \sum \frac{1}{(\rho_w - \rho_j)} \right],$$

where $m = 2$ for real ρ_w.

λ_{m-1} is of course, in general, a complex number, and can describe a transient response component of complex poles via the latter-time expression above. This expression for λ_{m-1} illustrates the difficulty of evaluation of transient response coefficients systems possessing roots of multiplicities greater than one. In the authors' notation then,

$$\lambda_{m-1} = \lim_{s \to \rho_w} \left[\frac{1}{1!} P' \right] = \lim_{s \to \rho_w} \left[\frac{1}{1!} PQ_1 \right] = \frac{1}{1!} \lambda_m Q_1 \Big|_{s=\rho},$$

since by previous notation,

$$\lim_{s \to \rho_w} kP = \lambda_m.$$

Let $\mathfrak{Q}_1$ denote $Q_1 \big|_{s=\rho_w}$; then the previous expression becomes

$$\lambda_{m-1} = \frac{1}{1!} \lambda_m \mathfrak{Q}_1.$$

Since

$$\mathfrak{Q}_l = \left[\sum_i \frac{1}{(\rho_w - z_i)^l} - \sum_{j \neq w} \frac{1}{(\rho_w - \rho_j)^l} \right],$$

then

$$\lambda_{m-2} = \frac{1}{2!} \lambda_m [\mathfrak{Q}_1^2 - \mathfrak{Q}_2]$$

$$\lambda_{m-3} = \frac{1}{3!} \lambda_m [\mathfrak{Q}_1^3 + 2\mathfrak{Q}_3 - 3\mathfrak{Q}_1 \mathfrak{Q}_2]$$

$$\lambda_{m-4} = \frac{1}{4!} \lambda_m [\mathfrak{Q}_1^4 + 3\mathfrak{Q}_2^2 - 6\mathfrak{Q}_4 + 8\mathfrak{Q}_3 \mathfrak{Q}_1 - 6\mathfrak{Q}_1 \mathfrak{Q}_2]$$

$$\lambda_{m-5} = \frac{1}{5!} \lambda_m [24\mathfrak{Q}_5 - 30\mathfrak{Q}_1 \mathfrak{Q}_4 - 20\mathfrak{Q}_2 \mathfrak{Q}_3 + 20\mathfrak{Q}_1^2 \mathfrak{Q}_3 + 15\mathfrak{Q}_1 \mathfrak{Q}_2^2 - 10\mathfrak{Q}_1^3 \mathfrak{Q}_2 + \mathfrak{Q}_1^5]$$

$$\lambda_{m-6} = \frac{1}{6!} \lambda_m [\mathfrak{Q}_1^6 - 120\mathfrak{Q}_6 + 144\mathfrak{Q}_1 \mathfrak{Q}_5 + 90\mathfrak{Q}_2 \mathfrak{Q}_4 + 40\mathfrak{Q}_3^2 - 90\mathfrak{Q}_1^2 \mathfrak{Q}_4 - 120\mathfrak{Q}_1 \mathfrak{Q}_2 \mathfrak{Q}_3 - 15\mathfrak{Q}_2^3 + 40\mathfrak{Q}_1^3 \mathfrak{Q}_3 + 45\mathfrak{Q}_1^2 \mathfrak{Q}_2^2 - 15\mathfrak{Q}_1^4 \mathfrak{Q}_2].$$

Another approach by which one may, without a knowledge of calculus, determine coefficients of high-order poles has been presented by Hazony and Riley [5] in a paper published subsequent to the preparation of this article.

Conclusion

Application of the preceding procedure to response functions having poles of high multiplicity permits calculation of the transient response coefficients associated with such poles with appreciably less effort than is required by other techniques.

Appendix

Example:

Given the following pole-zero configuration for some system excited with the unit step input shown in Fig. 1,

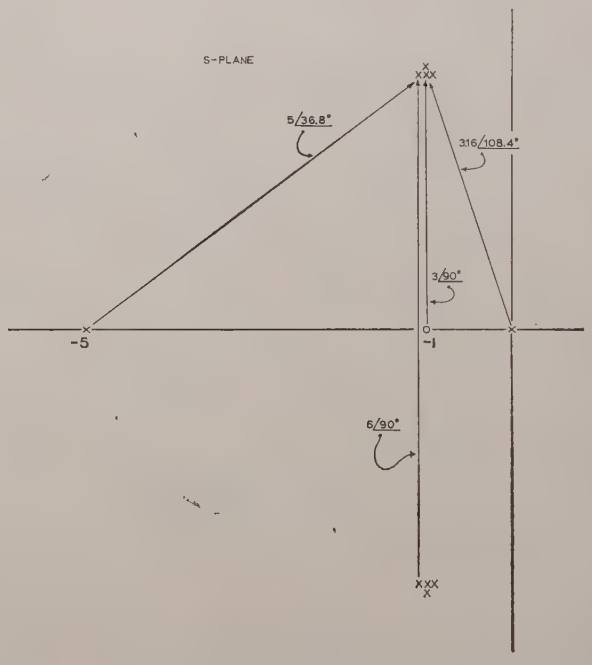


Fig. 1—Evaluation of transient response coefficients.

one calculates for the pole of multiplicity 4,

$$\mathfrak{Q}_1 = \left[\frac{1}{3/90^\circ} - \frac{1}{5/36.8^\circ} - \frac{1}{3.16/108.4} - \frac{4}{6/90^\circ} \right] = 0.753/94.6^\circ$$

$$\mathfrak{Q}_2 = \left[\frac{1}{9/180^\circ} - \frac{1}{25/73.6^\circ} - \frac{1}{9.98/216.8^\circ} - \frac{4}{36/180^\circ} \right] = 0.072/-17.45^\circ$$

$$\mathfrak{Q}_3 = \left[\frac{1}{27/270^\circ} - \frac{1}{125/110.4^\circ} - \frac{1}{31.6/325.2^\circ} - \frac{4}{216/270^\circ} \right] = 0.0245/161.0^\circ,$$

and obtains

$$\lambda_4 = \frac{k/90^\circ}{(5/36.8^\circ)(6^4/450)(3.16/108.4^\circ)}$$
$$= 0.0001464k/-55.2^\circ$$
$$\lambda_3 = \frac{1}{1!}\lambda_4 \mathfrak{R}_1 = \frac{1}{1!}(0.0001464k/-55.2^\circ)(0.753/94.6^\circ)$$
$$= 0.0001103k/39.4^\circ$$
$$\lambda_2 = \frac{1}{2!}\lambda_4[\mathfrak{R}_1^2 - \mathfrak{R}_2] = \tfrac{1}{2}(0.0001464k/-55.2^\circ)$$
$$\cdot[(0.753/94.6)^2 - (0.072/-17.95^\circ)]$$
$$= 0.0000461k/130.9^\circ$$
$$\lambda_1 = \frac{1}{3!}\lambda_4[2\mathfrak{R}_3 - 3\mathfrak{R}_1\mathfrak{R}_2 + \mathfrak{R}_1^3]$$
$$= \frac{1}{3!}(0.0001464k/-55.2^\circ)[2(0.0245/161.0^\circ)$$
$$- 3(0.072/-17.4^\circ)(0.753/94.6^\circ) + (0.753/94.6^\circ)^3]$$
$$= 0.00001358k/-144.9^\circ,$$

which yields the following portion of the time response due to the poles of multiplicity 4 at $-1 \pm j3$:

$$\cdots + 0.0000122kt^3\epsilon^{-t}\cos(3t - 55.2^\circ)$$
$$+ 0.0000368kt^2\epsilon^{-t}\cos(3t + 39.4^\circ)$$
$$+ 0.0000461kt\epsilon^{-t}\cos(3t + 130.9^\circ)$$
$$+ 0.00002716k\epsilon^{-t}\cos(3t - 144.9^\circ)$$
$$+ \cdots.$$

Bibliography

[1] G. J. Murphy, "Control Engineering," D. Van Nostrand Co., Inc., New York, N. Y., ch. 4; 1959.
[2] T. Jawor, "Using the root locus" *Control Engrg.*, vol. 6, pt. 2, pp. 119–122; November, 1959.
[3] W. R. Evans, "Control System Dynamics," McGraw-Hill Book Co., Inc., New York, N. Y.; 1954.
[4] W. Kaplan, "Advanced Calculus," Addison-Wesley Press, Cambridge, Mass., p. 19; 1952.
[5] D. Hazony and J. Riley, "Evaluating residues and coefficients of high order poles," IRE Trans. on Automatic Control, vol. AC-4, pp. 132–136; November, 1959.

Correction

Masanao Aoki, author of "On Optimal and Suboptimal Policies in the Choice of Control Forces for Final Value Systems," which appeared on pages 171–178 of the August, 1960, issue of these Transactions, has requested that the following Fig. 2 be substituted for the one which appeared on page 177 of the above.

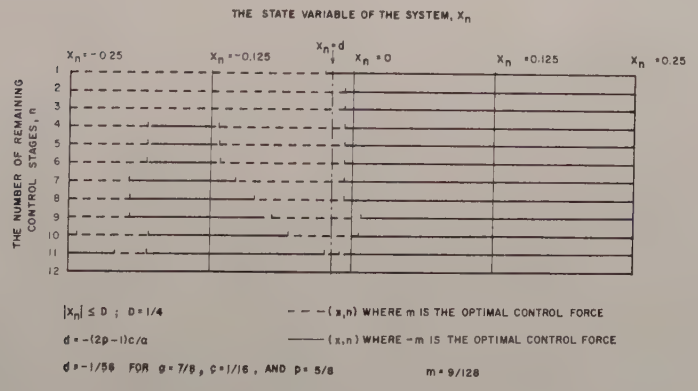


Fig. 2.

Correspondence

Discussion of "On Adaptive Control Systems"

The system described by Braun[1] is similar in philosophy to the system described by Mishkin and Haddad.[2] In the latter, an attempt was made to drive the error and its derivatives approximately to zero by the end of each control interval. In the system of Braun, it appears the error and its derivatives are to go instantly to zero at the start of each control interval and to remain approximately zero for the duration of the control interval. This is physically impossible in any real system. That is, with finite power we cannot achieve instantaneous change in the output.

Taking the Laplace transform of (26) yields the transfer function from error at the sampling instants to the additional actuator command,

$$E(s) = \Delta M(s)G(s), \tag{1}$$

or

$$\frac{\Delta M(s)}{E(s)} = \frac{1}{G(s)}. \tag{1a}$$

Eq. (1a) represents a physically unrealizable condition, for an arbitrary command signal. However, if we concede the approximate generation of impulses in ΔM, and $G(s)$ is of order 1, as in Braun's examples, (1) can be approximately realized. In this case, (29) is obtained by long division of G into E. If $G(s)$ is of order $q \geq 2$, Braun seems to suggest in (30) that we ignore the error and its first $q-2$ derivatives, apparently for the sake of compatible equations. This would result in poor control action, in which no control action was made on the error itself. Our driver might go parallel to the road, but not on it. A better alternative would be to include some sort of dynamics in (1) to achieve realizability, but necessarily abandoning the attempt to have immediate correspondence between input and output.

Both the systems of Mishkin and Haddad and of Braun are linear. If the time variation in the plant parameters is slow enough to be negligible, the stability analysis is relatively simple, using the techniques for the analysis of sampled-data systems.[3] Such an analysis has been made for the system of Mishkin and Haddad,[4] and the analysis of Braun's system proceeds in an identical fashion. A brief outline is given below.

The transfer function from error to actuator command will be taken as

$$\frac{E(s)}{\Delta M(s)} = \frac{1}{s^{q-1}G(s)}, \tag{2}$$

where q is the order of the plant. More precisely, (2) is the transfer function from the error just prior to the sampling instants, expressed as a MacLaurin series, to the additional actuator command applied at the sampling instants, expressed as a MacLaurin series. Some other physically realiazble transfer function could have been used instead of (2), perhaps with better control action, but in any case the analysis is the same. If we truncate to N terms, there results the block diagram of Fig. 1, with the various matrix quantities defined in Fig. 2.

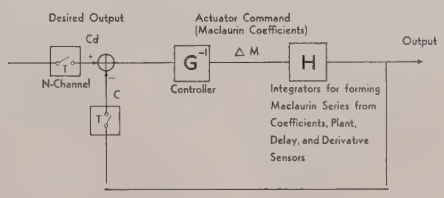


Fig. 1—Block diagram of system.

The slight delay in the transfer function H permits us to neglect instrument dynamics and yet avoid difficulties with discontinuous derivatives at sampling instants.

From the block diagram,

$$\Delta M = G^{-1}(C_d - C) \tag{3}$$

and

$$C = H(\Delta M). \tag{4}$$

Combining (3) and (4) gives

$$(1 + HG^{-1})C = HG^{-1}C_d. \tag{5}$$

For stability, then,

$$\text{Det}\,(1 + HG^{-1}) \neq 0 \quad \text{for} \quad |z| > 1. \tag{6}$$

Thus, the stability analysis solicited by Braun in the Conclusions is readily supplied, for the case of negligible parameter variation. A condition very much like (6) was determined[4] for the system of Mishkin and Haddad, and various examples worked. As might be expected in dealing with a linear feedback controller, the system was stable in some cases and unstable in others. The same would be true here.

If this idea is to be pursued further, it must first be modified for realizability. It would then be interesting to see it applied to a control problem worthy of such complex control means, such as a statically unstable airframe with pronounced flexibility problems. Solutions of (5) for $C(z)$ and testing for stability by the condition (6) would not be prohibitively difficult using a high-speed computer.

$$C_d \equiv \begin{bmatrix} C_d(Z) \\ \dot{C}_d(Z) \\ \vdots \\ C_d^{(N-1)}(Z) \end{bmatrix}, \quad \text{Desired Output (Sampled)} \qquad \Delta M \equiv \begin{bmatrix} \Delta M_{-1}(Z) \\ \Delta M_0(Z) \\ \vdots \\ \Delta M_{N-2}(Z) \end{bmatrix}, \quad \text{MacLaurin Coefficients of Actuator Command}$$

$$C \equiv \begin{bmatrix} C(Z) \\ \dot{C}(Z) \\ \vdots \\ C^{(N-1)}(Z) \end{bmatrix}, \quad \text{Output (Sampled)} \qquad G \equiv \begin{bmatrix} G_q & 0 & 0 & \cdots & 0 \\ G_{q+1} & G_q & 0 & & 0 \\ G_{q+2} & G_{q+1} & G_q & & \\ \vdots & & & \ddots & \\ G_{q+N-1} & & & & G_q \end{bmatrix}$$

$$\text{with } G_K \text{ given by } G(s) \equiv \sum_q^\infty \frac{G_K}{S^K}$$

$$H \equiv \begin{bmatrix} H_{-1} & H_0 & H_1 & \cdots & H_{N-2} \\ H_{-2} & H_{-1} & H_0 & & \\ H_{-3} & H_{-2} & H_{-1} & \ddots & \\ \vdots & & \ddots & \ddots & \\ H_{2-N} & & & & H_{-1} \end{bmatrix}, \qquad \text{with } H_K \equiv \mathcal{Z}\left[\frac{G(s)}{S^{1-K}} e^{-sT/a}\right]_{a \gg T}$$

Fig. 2—Definition of quantities in Fig. 1.

R. M. du Plessis
Autonetics Div.
North American Aviation, Inc.
Downey, Calif.

* Received by the PGAC, April 8, 1960.

[1] L. Braun, Jr., "On adaptive control systems," IRE Trans. on Automatic Control, vol. AC-4, pp. 30–42; November, 1959.

[2] E. Mishkin and R. A. Haddad, "Identification and command problems in adaptive systems," 1959 WESCON Convention Record, pt. 4, pp. 125–135.

[3] J. R. Ragazzini and G. F. Franklin, "Sampled Data Control Systems," McGraw-Hill Book Co., Inc., New York, N. Y.; 1958.

[4] R. M. du Plessis, "Application of the Z-Transformation to the Analysis of an Adaptive Control System," Internal Rept. Autonetics Div., North American Aviation, Inc., Downey, Calif.

Author's Comment[5]

Mr. du Plessis has made a number of interesting and constructive comments regarding the subject paper by this author. He points out the undesirable aspects of impulse drives—correctly, of course. He further indicates that, in (30) of the subject paper, the value of the error and its first $q-2$ derivatives are ignored, "for the sake of compatible equations." This latter point is not quite correct. If $G(s)$ is of order q, and $\Delta M(s)$ is a linear combination of an impulse and the first m integrals of an impulse, then it may be shown, using the initial-value theorem, that $e(0+)$ and the first $q-2$ derivatives of $e(t)$—all evaluated at $t=0+$—are identically zero. If $\Delta M(s)$ contains higher-order impulses, fewer than $q-2$ derivatives of $e(t)$ are zero.

Mr. du Plessis points out the similarity of the philosophy of the author's approach to that taken by Mishkin and Haddad. This author has worked in the same research group with Mishkin and Haddad for several years. The philosophy in the subject paper and in the Mishkin-Haddad paper is an outgrowth of a mutual effort in this area. The Mishkin-Haddad system is clearly superior to the one proposed by this author—in fact, the Mishkin-Haddad system is a result of efforts to overcome the practical difficulties in the earlier system.

In the opinion of the author, the contribution (if there is any) of the paper is in the philosophy developed; the particular application having been introduced only for illustrative purposes.

Mr. du Plessis has made an interesting application of sampled-data theory in the stability analysis he suggests.

Ludwig Braun, Jr.
Polytechnic Inst. of Brooklyn
Brooklyn, N. Y.

[5] Received by PGAC, July 8, 1960.

Two Digital Computer Programs for Use with Multirate Sampled-Data System Analysis*

Two digital computer programs were written last year at North American Aviation, Inc., in connection with the analysis of multirate sampled-data systems, which might be of interest to people engaged in similar analysis work. The systems under consideration were more or less like the block diagram of Fig. 1.

These programs were entitled 1) Multirate Z-Transform Program and 2) Multirate Z-Frequency-Response Program. They were programmed in the FORTRAN system on the IBM 709 by J. C. Long of the Missile Division of North American Aviation, Inc.

* Received by the PGAC, July 21, 1960.

Multirate Z-Transform Program

The purpose of this program is to transform expressions in $z_n \equiv e^{sT_n}$, representing a time function sampled "fast" at T/n, into an expression in $z \equiv e^{sT}$, representing the same time function sampled "slow" at T. In order to make this transformation, a table suitable for look-up operations in a computer was derived. This table was derived by the formula given by Ragazzini and Franklin,[1]

$$Z[G(z_n)] = \frac{1}{2\pi j}\int_{\Gamma} \frac{G(z_n)}{z_n}\,\frac{dz_n}{(1-z_n{}^n z^{-1})} \qquad \text{poles of } \frac{G(z_n)}{z_n}. \tag{1}$$

Some fairly simple transform pairs resulted, as shown in Table I.

The method of computation is to break up the expression to be transformed into partial fractions of the type in Table I, to transform according to Table I, and to recombine, and factor the result. This process is illustrated in Fig. 2.

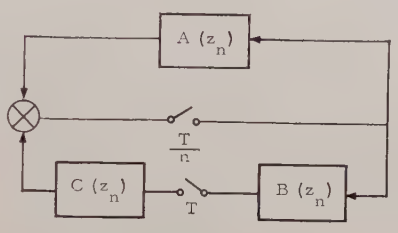


Fig. 1.

Multirate Z-Frequency Response Program

The purpose of this program is to take expressions in $z_n \equiv e^{sT/n}$, representing a time function "fast" sampled at T/n, and to obtain directly the frequency response, or mapping of the unit circle, of $Z[G(z_n)]$, "slow" sampled at T. The basis of this program is a formula appearing in Jury:[2]

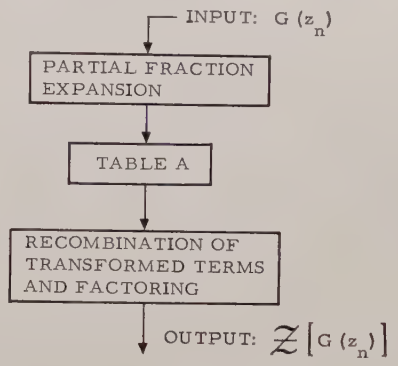


Fig. 2.

TABLE I
Multirate Z-Transform Pairs for Sampled Data System Analysis

$G(z_n)$	$Z[G(z_n)]$
$kG(z_n)$	$kZ[G(z_n)]$
$z_n{}^{-p},\ p \neq kn$	0
$z_n{}^{-p},\ p = kn$	z^{-k}
$\dfrac{1}{(z_n - a)}$	$\dfrac{a^{n-1}}{z - a^n}$
$\dfrac{1}{(z_n - a)^2}$	$\dfrac{a^{2n-2} + (n-1)a^{n-2}z}{(z - a^n)^2}$
$\dfrac{1}{(z_n - a)^3}$	$\dfrac{2a^{3n-3} + (n+4)(n-1)a^{2n-3}z + (n-1)(n-2)a^{n-3}z^2}{2!(z - a^n)^3}$
$\dfrac{1}{(z_n - a)^4}$	$\dfrac{1}{3!(z - a^n)^4}[6a^{4n-4} + (n-1)(n^2+7n+18)a^{3n-4}z + (n-1)(4n^2+4n-18)a^{2n-4}z^2 + (n-1)(n-2)(n-3)a^{n-4}z^3]$
$\dfrac{1}{z_n{}^2 + 2az_n + a^2 + b^2}$	$\dfrac{b(a^2+b^2)^{n-1} + (a^2+b^2)^{(n-1)/2}\sin[(n-1)\theta]z}{b[z^2 - 2(a^2+b^2)^{(n)/2}\cos(n\theta)z + (a^2+b^2)^n]}$ $\theta \equiv \tan^{-1}\dfrac{b}{a}$
$\dfrac{z_n}{z_n{}^2 + 2az_n + a^2 + b^2}$	$\dfrac{(a^2+b^2)^{(n)/2}\sin(n\theta)z}{b[z^2 - 2(a^2+b^2)^{(n)/2}\cos(n\theta)z + (a^2+b^2)^n]}$ $\theta \equiv \tan^{-1}\dfrac{b}{-a}$
$F(z_n{}^n)G(z_n)$	$F(z)Z[G(z_n)]$

(factors of the form $F(z_n{}^n)$ can be excluded, for convenience, from data input to computer)

[1] J. R. Ragazzini and G. F. Franklin, "Sampled Data Control Systems," McGraw-Hill Book Co., Inc., New York, N. Y., p. 227; 1958.

[2] E. I. Jury, "Sampled Data Control Systems," John Wylie and Sons, Inc., New York, N. Y., p. 75; 1958.

$$Z[G(z_n)] = \frac{1}{n}\sum_{k=1}^{n} G(z_n e^{j2\pi k/n}). \quad (2)$$

The method of calculation is illustrated in Fig. 3. The input to the program is an expression $G(z_n)$ and the desired frequency points to be computed. These frequencies range from

$$0 \quad \text{to} \quad \frac{1}{2}\,\frac{2\pi}{T},$$

because higher frequencies lie on the same curve, or its conjugate.

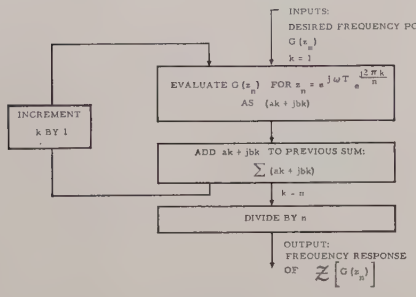


Fig. 3.

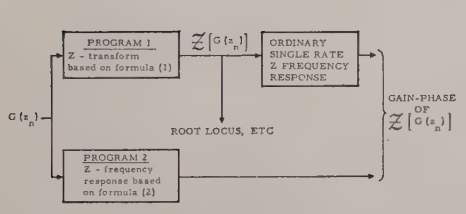


Fig. 4.

In actual practice, the second program has proved more convenient to use, because the output is already in a form for any desired frequency-domain stability plot, whereas the output of the first requires further analysis to determine stability (root locus). The relationship between these two programs is shown in Fig. 4.

R. M. du Plessis
Autonetics
North American Aviation, Inc.
Downey, Calif.

Short-Time Stability*

The concept of stability plays a dominant role in control systems. However, the classical definitions—for example, "A linear varying-parameter system is defined to be stable if and only if every bounded input produces a bounded output"[1]—suffer from some important shortcomings. One of them is that knowing that an output will be bounded is generally not sufficient. The exact value of the bound must also be known. In addition, the time interval of interest is often only of finite length. A case in point is that of a missile whose operation lasts only a finite time and whose variables must be restrained to certain critical values in this time interval. To overcome some of these difficulties, the following definition is proposed:

Definition—A linear system with input $x(t)$ and output $y(t)$ is said to be short-time stable in the interval $0 \le t \le T$ with respect to ϵ and C if $|x(t)| \le \epsilon$ implies $|y(t)| \le C$ in the interval $0 \le t \le T$.

A similar definition for unperturbed (no forcing function) systems is given by Chzhan-sy-in.[2] To illustrate this concept of stability, a theorem and some examples follow.

Theorem—The necessary and sufficient condition for a linear time invariant system with impulsive response $w(t)$ to be short-time stable over $0 \le t \le T$ with respect to ϵ and C is

$$\int_0^T |w(t)|\,dt \le C/\epsilon. \quad (1)$$

The sufficiency follows from the set of inequalities,

$$|y(t)| \le \int_0^t |w(t')|\,|x(t-t')|\,dt' \le \epsilon \int_0^t |w(t')|\,dt' \le C \quad (2)$$

and

$$\int_0^t |w(t')|\,dt' \le \int_0^T |w(t')|\,dt'. \quad (3)$$

To prove necessity requires that, when

$$\int_0^T |w(t)|\,dt > C/\epsilon,$$

an input with $|x(t)| \le \epsilon$ can be found that causes an output $|y(t)| > C$ for some t in $0 \le t \le T$. The required input is

$$x(T-t') = +\epsilon \quad \text{for} \quad w(t') > 0,$$
$$x(T-t') = -\epsilon \quad \text{for} \quad w(t') < 0, \quad (4)$$

for then,

$$y(T) = \int_0^T w(t')x(T-t')dt' = \epsilon \int_0^T |w(t')|\,dt' > C. \quad (5)$$

For an example, consider $w(t) = Ke^t$, which is unstable in the classic sense regardless of the value of K, excluding, of course, $K = 0$. However, it will be short-time stable if K satisfies the inequality,

$$|K| \int_0^T e^t dt \le C/\epsilon. \quad (6)$$

The required value of K is easily seen to be,

$$|K| \le \frac{C}{\epsilon(e^T - 1)} \quad (7)$$

Note that the same system may be considered stable or unstable depending on the particular choice of T, C, and ϵ. Consider next $w(t) = e^{-t} - e^{-2t}$. This system is classically stable. However it can become short time unstable if T is too large since,

$$\int_0^T |w(t)|\,dt = \tfrac{1}{2}(e^{-T} - 1)^2 > C/\epsilon, \quad (8)$$

for T sufficiently large and, say, $C/\epsilon \le \frac{1}{4}$.

It is hoped that this short note will stimulate interest in this concept of stability and be of some use to control system engineers.

Acknowledgment

Appreciation is extended to Prof. H. Hochstadt of the mathematics department for his useful suggestions.

Peter Dorato
Polytechnic. Inst. of Brooklyn
Brooklyn, N. Y.

Sampling Schemes in Sampled-Data Control Systems*†

Introduction

In the early development of the sampled-data field, it was assumed that the sampling pattern is periodic with fixed period T, that the pulse width is of negligible duration, and that these short pulses are represented by impulses. Based on this representation, several analysis methods were developed, among them the z-transform method[1]–[3], the difference equation [4]–[6], and the state-variable approach [6], [7]. Following recent extension and research in this field, these preliminary assumptions have been relaxed to a large extent, so that wider application and more general methods of analysis could be developed. In the discussion of the first category on the fixed sampling pattern, it can be easily observed that the z-transform method with its various extensions is applicable to deal with all the indicated cases.

For the second category, which includes varying sampling times and varying pulse width duration, the z-transform method is applied only in approximation cases, but the difference equation approach has more promise. Similarly, the latter is applicable to random sampling. It is assumed that in the first category the system operation is linear, with time-invariant and time-varying parameters. In the second category, the system operation is essentially nonlinear; and in the third category, the system description is linear with randomly varying parameters.

Sampling Schemes

To illustrate clearly the distinction between the various operating conditions of the sampled-data systems, the sampling schemes are divided and discussed in the following three categories:

* Received by the PGAC, August 5, 1960.
[1] L. A. Zadeh, "On stability of linear varying-parameter systems," *J. Appl. Phys.*, vol. 22, pp. 402–405; April, 1951.

[2] Chzhan-sy-in, "Stability of motion during a finite time interval," *J. Appl. Math. and Mechanics*, vol. 23, pp. 333–344; 1959.

* Received by the PGAC, October 17, 1960.
† Based on a lecture delivered on April 29, 1960, at the University-Industry Colloquium, University of California, Berkeley.

1) *Fixed sampling pattern:* In this part, only the fixed or predetermined pattern is included which characterizes the operation of the sampler (or the flow of information). This is divided into the following:

a) Periodic sampling with fixed period T and negligible pulse width duration (impulses), shown in Fig. 1.
b) Staggered sampling with same period, shown in Fig. 2.
c) Multi-rate sampling (Fig. 3).
d) Multiple samplers operation (Fig. 4).
e) Cyclic-rate sampling (periodically time-varying sampling rate) (Fig. 5).
f) Skip-sampling, also called time-quantitized aperiodic sampling [1] (Fig. 6): In this case one or more samples of a signal may be periodically or generally omitted. The samples occur only at integral multiples of some minimal period, but do not occur all the time.
g) Slowly varying sampling rate (Fig. 7): The change per sample of the period is small compared to the period [1].
h) Piecewise constant rate sampling [1] (Fig. 8).
i) Finite pulse width sampling (Fig. 9).
j) Periodically varying sampling rate and pulse width (Fig. 10).
k) General aperiodic sampling of varying pulse-width (Fig. 11).
l) Signal sampling occurring at various time gates γ_0, γ_1, γ_2, as shown in Fig. 12, are sometimes used for the purpose of better filtering schemes of signal and noise [7].

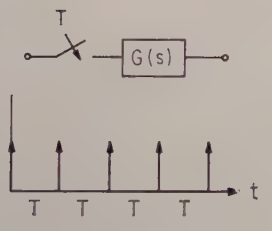


Fig. 1—Periodic sampling with fixed period "T."

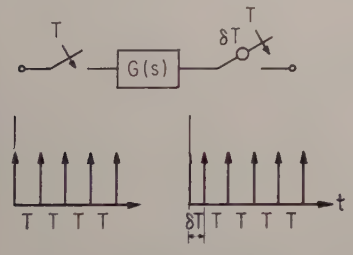


Fig. 2—Staggered sampling.

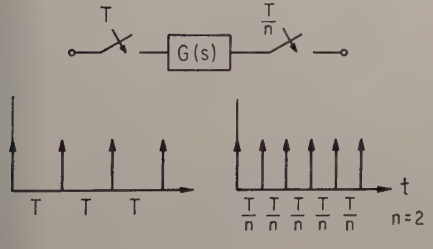


Fig. 3—Multi-rate sampling.

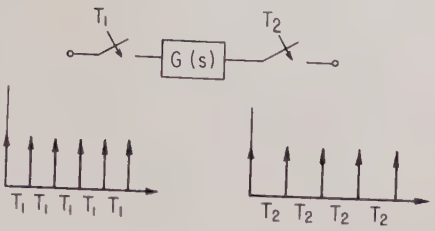


Fig. 4—Multiple samplers operation.

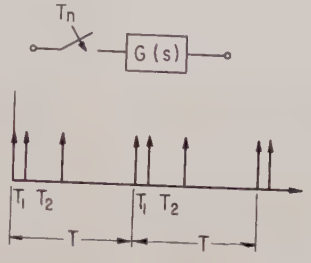


Fig. 5—Cyclic-rate sampling.

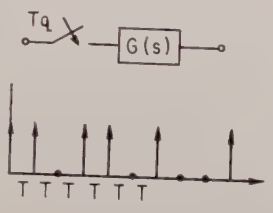


Fig. 6—Time-quantized sampling.

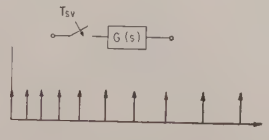


Fig. 7—Slowly varying sampling rate.

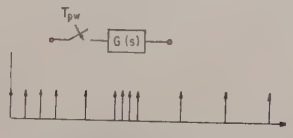


Fig. 8—Piecewise constant-rate sampling.

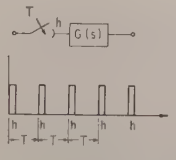


Fig. 9—Finite pulse-width sampling.

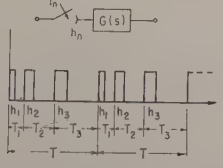


Fig. 10—Periodically varying sampling rate and pulse width.

2) *Signal-dependent sampling:* In this sampling scheme, the operation of the sampler is not *a priori* fixed, but is dependent on the signals flowing in the system, usually the error signal. The basic operation of such a sampler in a linear system is a nonlinear one. This is unlike the first sampling pattern which constitutes a linear operation, provided the system is linear. The signal-dependent schemes can be realized in a variety of ways; however, in recent literature only two basic schemes are discussed. They are enumerated below:

a) Variable-phase sampling [1]: In this scheme, the nth sample time occurs before the time $(n+\frac{1}{2})T$, and after $(n-\frac{1}{2})T$, where T is an average sampling period and is constant independent of "n," *i.e.*, $(n-\frac{1}{2})T<t_n<(n+\frac{1}{2})T$ (where t_n is the time of nth sampling pulse). A typical form of t_n is $t_n=T[n+p(t_n)]$, where $p(t)$ is a function of the system signals. A schematic diagram of such a sampling pattern is shown in Fig. 13.

b) Variable pulse-width sampling [4]: This is also referred to as pulse-width modulation and can be accomplished in one of the following three forms: 1) lead-type,

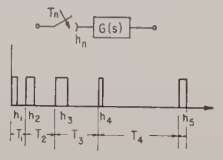


Fig. 11—General aperiodic sampling.

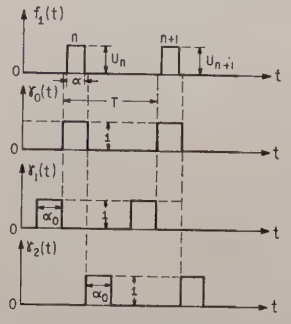


Fig. 12—Train of predetermined pulses of $f_1(t)$ and the various gate pulses.

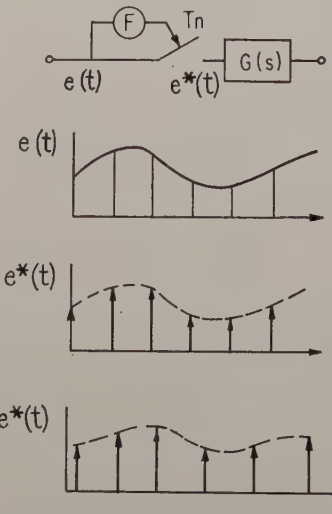


Fig. 13—Variable-phase sampling.

2) lag-type, and 3) lag-delay-integrator type. These various types of pulse-width modulation are shown in Fig. 14. The operation of a magnetic amplifier in a control system is basically of the third type, pulse-width modulation.

c) Zero-signal sampling [6]: In efficient transmission of information and for maximum bandwidth, it is sometimes necessary to sample the time at zero values of the signal, as shown in Figs. 15(a) and 15(b). In this case, information theory plays an important role in the analysis.

3) *Random sampling schemes* [1], [6], [7]: In this case, the operation of the sampler occurs at random (Fig. 16), either in virtue of "misses" in information or "jitter" in a sampling mechanism, or is purposely introduced, for reasons of economy in time-sharing of a digital computer among several processes or for reducing susceptibility to jamming or interference. Sampled-data systems with random sampling and randomly varying parameter are described by difference equations with randomly varying coefficients [1]. Stability study of such equations is being pursued recently. Further study of such systems is being based on the statistical properties of the random input and sampling [9].

Discussion and Use of the Various Sampling Schemes

The variety of the sampling schemes discussed earlier indicates the extension of theory and application from the orthodox sampled-data system when originally introduced. Some of these schemes have practical use, either because of inherent operation of the sampler, or because of purposeful introduction of such schemes for reasons of compensation, economy and optimization procedures. For the analysis of linear systems with fixed sampling schemes, the z transform and the time-varying z transform are applicable, and the problems of stability, response to various inputs, and compensation can be easily handled with the existing theory.

Regarding the signal-dependent sampling schemes, the linear theory is no longer adequate except for linearization approximation; however, much of the recent advance in nonlinear theory of difference equations is applicable, including Lyapunov's second method for determining asymptotic stability in the case of pulse-width modulated systems [4].

Similarly, much activity is being directed toward study of discrete systems with randomly varying parameters and sampling schemes, which gives the answer to several of the problems connected with the third category.

Conclusions

In this short discussion, the aim of the writer has been to set forth the problems associated with the various sampling schemes in the field of sampled-data and digital control systems, and to indicate specifically the important mathematical tools available for dealing with these arising situations. From an engineering point of view, one should also look into the various approximations which simplify the solution of the problem considerably, without any significant deviation from the exact solution.

E. I. Jury
Dept. of Elec. Engrg.
University of California
Berkeley, Calif.

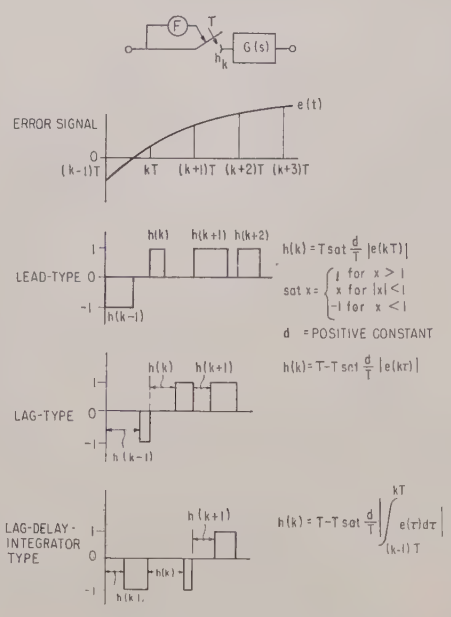


Fig. 14—Variable pulse-width sampling.

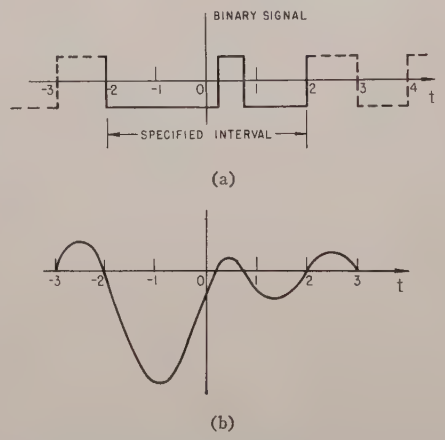


Fig. 15—Signal-dependent sampling scheme (sampling time occurring at zeros of signal) [8]. (a) Binary signal. (b) Band-limited signal having the specified zero crossing.

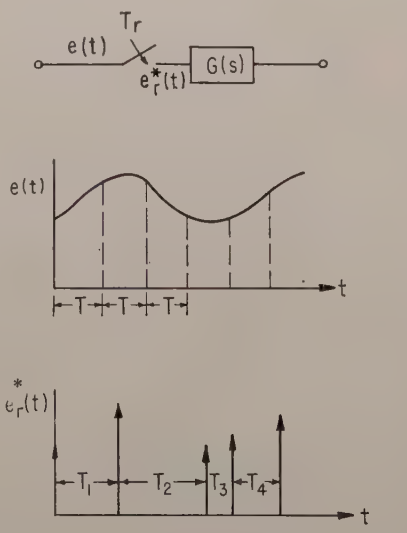


Fig. 16—Random-sampling scheme.

References

[1] R. E. Hufnagel, "Analysis of Aperiodically-Sampled-Data Feedback Control Systems," Ph.D. dissertation, College of Engrg., Cornell University, Ithaca, N. Y.; June, 1959.
[2] E. I. Jury, "Recent Advances in the Field of Sampled-Data and Digital Control Systems," presented at the IFAC Congress, Moscow, USSR; June, 1960. To be published by Butterworth Scientific Publications, London, Eng.
[3] Y. Z. Tsypkin, "Theory of Pulse Systems," State Press for Physical and Mathematical Literature, Moscow, USSR; 1959. (In Russian.)
[4] T. T. Kadota, "Analysis of Non-Linear Sampled-Data Systems with Pulse-Width Modulation," Ph.D. dissertation, Dept. of Elec. Engrg., University of California, Berkeley; April, 1960.
[5] B. Friedland, "Sampled-Data Control Systems Containing Periodically Varying Members," Dept. of Elec. Engrg., Columbia University, New York, N. Y., Tech. Rept. T-39/B; November 12, 1959.
[6] R. E. Kalman and J. E. Bertram, "A unified approach to the theory of sampling systems," *J. Franklin Inst.*, vol. 267, pp. 405–436; May, 1959.
[7] F. M. Kilin, "Some Problems in the Theory of Pulse Systems with Time Gates," IFAC Congress preprints, Moscow, U.S.S.R. vol. 3; June–July, 1960. To be published by Butterworths Scientific Publications, London, Eng.
[8] F. E. Bond and C. R. Cahn, "On sampling the zeros of bandwidth limited signals," IRE Transactions on Information Theory, vol. IT-4, pp. 110–113; September, 1958.
[9] A. R. Bergen, "On the Statistical Design of Linear Random Sampling," presented at the IFAC Congress, Moscow, USSR, June, 1960.

Notes on the Stability Criterion for Linear Discrete Systems*

It is known that linear time-invariant discrete systems can be described by constant coefficient linear difference equations. One of the problems in the analysis of such systems is the test for stability. These tests involve both graphical procedures such as Nyquist locus, Bode diagrams and the root-locus, and analytical methods such as Schur-Cohn[1] or Routh-Hurwitz criteria. Because of the high-order determinants to be evaluated using the present form of the Schur-Cohn criterion, many authors have used the bilinear transformation which maps the inside of the unit circle in the $z=e^{Ts}$ plane into the left half of the w plane and then applied the Routh-Hurwitz criterion. This transformation involves algebraic manipulation which for higher-order systems becomes complicated.

In this discussion, it is shown that the evaluation of the Schur-Cohn[1] determinants can be simplified considerably, so that the manipulations involved in testing for the zeros of a polynomial are comparable to those using the "transformed" Routh-Hurwitz criterion, thus avoiding the bilinear

* Received by the PGAC, October 20, 1960.
[1] M. Marden, "The Geometry of the Zeros of a Polynomial in a Complex Variable," American Mathematical Society, New York, N. Y.; 1949.

transformation. This simplification is based on: 1) transforming the original Schur-Cohn $2k$-order matrix Δ_k by a unitary transformation to a new matrix whose symmetrical structure is utilized to reduce its determinant, $|\Delta_k|$, to a product of two determinants of matrices of order k; and, 2) exploiting the symmetry properties of the k-order matrices to reduce the evaluation of the sign of $|\Delta_k|$ to the identification of certain terms in the expansion of one of the product determinants. Thus, for an n-order polynomial in z, the test for stability involves evaluating determinants up to order n, a situation generally similar to the Routh-Hurwitz criterion.

Schur-Cohn Criterion

If for the polynomial

$$F(z) = a_0 + a_1z^2 + a_2z^2 + \cdots a_nz^n, \quad (1)$$

all the determinants of the matrices

$$\Delta_k = \begin{bmatrix} a_0 & 0 & 0 & \cdots 0 & a_n & a_{n-1} & \cdots a_{n-k+1} \\ a_1 & a_0 & 0 & \cdots 0 & 0 & a_n & \cdots a_{n-k+2} \\ \cdot & \cdot & \cdot & \cdots\cdot & \cdot & \cdot & \cdots\cdot \\ a_{k-1} & a_{k-2} & a_{k-3} & \cdots a_0 & 0 & 0 & \cdots a_n \\ \bar{a}_n & 0 & 0 & \cdots 0 & \bar{a}_0 & \bar{a}_1 & \cdots \bar{a}_{k-1} \\ \bar{a}_{n-1} & \bar{a}_n & 0 & \cdots 0 & 0 & \bar{a}_0 & \cdots \bar{a}_{k-2} \\ \cdot & \cdot & \cdot & \cdots\cdot & \cdot & \cdot & \cdots\cdot \\ \bar{a}_{n-k+1} & \bar{a}_{n-k+2} & \bar{a}_{n-k+3} & \cdots \bar{a}_n & 0 & 0 & \cdots \bar{a}_0 \end{bmatrix} \quad (2)$$

$k = 1, 2, \cdots, n$

$\bar{a}_k$ = complex conjugate of a_k,

are different from zero, then $F(z)$ has no zeros on the circle $|z| = 1$ and μ zeros in this circle, μ being the number of variations in sign in the sequence 1, $|\Delta_1|$, $|\Delta_2|$, $\cdots$, $|\Delta_n|$.

For a system of order n to be stable, all the n zeros of its characteristic nth-order equation must lie within the unit circle, *i.e.*, the sequence 1, $|\Delta_1|$, $|\Delta_2|$, $\cdots$, $|\Delta_n|$ must have n variations in sign. The stability can therefore be expressed by the constraints:[2]

$$\begin{aligned} |\Delta_k| &< 0, \quad k \text{ odd}, \\ |\Delta_k| &> 0, \quad k \text{ even}. \end{aligned} \quad (3)$$

For a discrete or a sampled-data system, all the coefficients of the characteristic equation are real. Hence, the conjugate sign on (2) is superfluous.

As noticed from (2), the highest-order determinant $|\Delta_n|$ is of order $2n$, while the characteristic equation is of order n. This fact was discouraging in using this criterion heretofore for higher-order sampled-data systems, the easy alternative being to transform to the w plane.

Simplification of the Stability Constraint Equation[3]

Let $\mathcal{J}_k$ be the k-order identity matrix, $\mathcal{J}_k^+$ the k-order permutation matrix

$$\mathcal{J}_k^+ = \begin{bmatrix} 0 & 0 \cdots 0 & 1 \\ 0 & 0 \cdots 1 & 0 \\ \cdot & \cdots\cdot & \cdot \\ 0 & 1 \cdots 0 & 0 \\ 1 & 0 \cdots 0 & 0 \end{bmatrix},$$

and U_k the $2k$-order unitary matrix

$$U_k = \begin{bmatrix} \mathcal{J}_k & 0 \\ 0 & \mathcal{J}_k^+ \end{bmatrix}. \quad \text{(Note that } U_k^{-1} = U_k\text{)}.$$

Let $\Lambda_k = U_k^{-1}\Delta_k^T U_k$, where the superscript T denotes transpose. Then,

$$|\Lambda_k| = |\Delta_k^T| = |\Delta_k|$$

and

$$\Lambda_k = \begin{bmatrix} X_k & Y_k \\ Y_k & X_k \end{bmatrix},$$

where

$$X_k = \begin{bmatrix} a_0 & a_1 & a_2 \cdots a_{k-1} \\ 0 & a_0 & a_1 \cdots a_{k-2} \\ \cdot & \cdot & \cdot\cdots\cdot \\ 0 & 0 & 0 \cdots a_1 \\ 0 & 0 & 0 \cdots a_0 \end{bmatrix},$$

and

$$Y_k = \begin{bmatrix} a_{n-k+1} \cdots a_{n-1} & a_n \\ a_{n-k+2} \cdots a_n & 0 \\ \cdot \quad\cdots\quad \cdot & \cdot \\ a_{n-1} \quad \cdots 0 & 0 \\ a_n \quad \cdots 0 & 0 \end{bmatrix}.$$

Hence,

$$\begin{aligned} |\Delta_k| &= |\Lambda_k| = \begin{vmatrix} X_k & Y_k \\ Y_k & X_k \end{vmatrix} \\ &= \begin{vmatrix} X_k + Y_k & Y_k + X_k \\ Y_k & X_k \end{vmatrix} \\ &= \begin{vmatrix} X_k + Y_k & 0 \\ Y_k & X_k - Y_k \end{vmatrix} \\ &= |X_k + Y_k|\,|X_k - Y_k|, \end{aligned}$$

the product of two k-order determinants which is considerably easier to evaluate than the direct evaluation of the $2k$-order determinant $|\Delta_k|$.

The Symmetrical Properties of $|X_k+Y_k|\,|X_k-Y_k|$

Now $|X_k+Y_k|$ is a homogeneous polynomial of dimension k in the variables $a_1, \cdots, a_n$. The polynomial $|X_k-Y_k|$ is identical to the polynomial $|X_k+Y_k|$ except for a change of sign of those monomial terms which have an odd number of elements from Y_k, *i.e.*,

$$\begin{aligned} |X_k + Y_k| &= A_k + B_k, \\ |X_k - Y_k| &= A_k - B_k, \end{aligned}$$

where $A_k(B_k)$ is the sum of all monomial terms which do not change (do change) sign where Y_k is replaced by $-Y_k$ in $|X_k+Y_k|$.

To identify A_k and B_k:

1) Let all the a_i's in the matrix Y_k in (9) be denoted by b_i's; then expand the determinant $|X_k+Y_k|$ in terms of a_i and b_i.

2) After expansion, examine every term which is a product of a_i's and b_i's; if it contains an *even* number of b_i's, then it is assigned to A_k; otherwise, assign the term to B_k.

3) After collecting the terms of A_k and B_k, replace all the b_i's by the a_i's.

Hence,

$$|\Delta_k| = (A_k + B_k)(A_k - B_k) = A_k^2 - B_k^2,$$

and for the stability this reduces to $|A_k| \gtrless |B_k|$ depending on k. The application of the Schur-Cohn criterion now reduces to the evaluation of determinants up to order n only for nth-order polynomials. The transformation to the w plane in order to use the Routh-Hurwitz criterion is no longer necessary and, in some cases, is more involved than the procedure outlined above.

Examples

In the following, stability criteria for systems up to a fourth-order system are established using the above techniques.

A. Second-Order System, $n=2$

$$F(z) = a_0 + a_1z + a_2z^2.$$

For $k=1$, $X_1=a_0$, $Y_1=a_2$. Thus,

$$\begin{aligned} X_1 + Y_1 &= a_0 + a_2, \\ |\Delta_1| &= |X_1 + Y_1|\,|X_1 - Y_1|. \end{aligned}$$

In this case, $A_1=a_0$, $B_1=a_2$.
For stability,

$$\begin{aligned} |\Delta_1| &< 0, \quad \text{or} \quad A_1^2 - B_1^2 < 0, \\ \textit{i.e.,}\ |A_1| &< |B_1| \text{ or } |a_0| < |a_2|. \end{aligned}$$

For $k=2$,

$$X_2 = \begin{vmatrix} a_0 & a_1 \\ 0 & a_0 \end{vmatrix}$$

$$Y_2 = \begin{vmatrix} a_1 & a_2 \\ a_2 & 0 \end{vmatrix} = \begin{vmatrix} b_1 & b_2 \\ b_2 & 0 \end{vmatrix}.$$

Following our procedure for identifying A_2 and B_2, we replaced all the a's of Y_2 by b's.

$$|\Delta_2| = |X_2 + Y_2|\,|X_2 - Y_2|.$$

Now,

$$\begin{aligned} |X_2 + Y_2| &= \begin{vmatrix} a_0 + b_1 & a_1 + b_2 \\ +b_2 & a_0 \end{vmatrix} \\ &= a_0^2 + a_0b_1 - a_1b_2 - b_2^2. \end{aligned}$$

Thus,

$$\begin{aligned} A_2 &= a_0^2 - b_2^2 = a_0^2 - a_2^2, \\ B_2 &= a_0b_1 - a_1b_2 = a_0a_1 - a_1a_2. \end{aligned}$$

[2] E. I. Jury, "Sampled-Data Control Systems," John Wiley and Sons, Inc., New York, N. Y.; 1958.
[3] B. H. Bharucha, "Analysis of Integral-Square Error in Sampled-Data Control Systems," Electronics Res. Lab., University of California, Berkeley, Issue No. 206, Series No. 60; 1958.

For stability,

$$|\Delta_2| > 0, \quad \text{or } |A_2| > |B_2|,$$
$$i.e., \ |a_0 + a_2| > |a_1|.$$

B. Third-Order System, $n=3$

$$F(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3.$$
$$k = 1, \quad X_1 = a_0, \quad Y_1 = a_3,$$
$$X_1 + Y_1 = a_0 + a_3$$
$$A_1 = a_0, \quad B_1 = a_3.$$

For $|\Delta_1| < 0$, $|a_0| < |a_3|$.
For $k=2$,

$$X_2 = \begin{bmatrix} a_0 & a_1 \\ 0 & a_0 \end{bmatrix} Y_2 = \begin{bmatrix} a_2 & a_3 \\ a_3 & 0 \end{bmatrix}$$
$$|X_2 + Y_2| = \begin{vmatrix} a_0 + a_2 & a_1 + a_3 \\ a_3 & a_0 \end{vmatrix}$$
$$= a_0^2 + a_0 a_2 - a_1 a_3 - a_3^2$$
$$A_2 = a_0^2 - a_3^2, \quad B_2 = a_0 a_2 - a_1 a_3.$$

For

$$|\Delta_2| > 0, \quad |A_2| > |B_2|, \text{ or}$$
$$|a_0^2 - a_3^2| > |a_0 a_2 - a_1 a_3|.$$

For $k=3$,

$$X_3 = \begin{bmatrix} a_0 & a_1 & a_2 \\ 0 & a_0 & a_1 \\ 0 & 0 & a_0 \end{bmatrix}, \quad Y_3 = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_2 & a_3 & 0 \\ a_3 & 0 & 0 \end{bmatrix},$$
$$A_3 = a_0^3 + a_0 a_1 a_3 - a_0 a_2^2 + a_1 a_2 a_3 - a_2 a_3^2 - a_0 a_3^2$$
$$B_3 = a_3 a_0^2 + a_1 a_0^2 + a_1^2 a_3 - a_0 a_1 a_2 - a_0 a_2 a_3 - a_3^3.$$

Thus, for $|\Delta_3| < 0$, or the system to be stable,

$$|A_3| < |B_3|.$$

C. Fourth-Order System, $n=4$

The conditions for stability for $n=4$ are given below.

$$F(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4.$$
$k = 1$
$$A_1 = a_0, \quad B_1 = a_4.$$
$k = 2$
$$A_2 = a_0^2 = a_4^2,$$
$$B_2 = a_0 a_3 - a_1 a_4.$$
$k = 3$
$$A_3 = a_0^3 + a_0 a_2 a_4 + a_1 a_3 a_4 - a_0 a_4^2 - a_2 a_4^2 - a_0 a_3^2,$$
$$B_3 = a_0^2 a_4 + a_0^2 a_2 + a_1^2 a_4 - a_0 a_2 a_4 - a_4^3 - a_0 a_1 a_3.$$
$k = 4$
$$A_4 = a_0^4 - a_0^2 a_2^2 - a_0^2 a_3^2 - 2a_0^2 a_4^2 - a_1^2 a_4^2 - a_2^2 a_4^2 + a_0^2 a_1 a_3 + 2a_0 a_2^2 a_4 - a_0 a_3^2 a_4 - a_0 a_1^2 a_4 - a_1^2 a_2 a_4 + a_1 a_3 a_4^2 - a_0 a_2 a_3^2 + a_0 a_1 a_2 a_3 + 2a_{01} a_3 a_4 + a_1 a_2 a_3 a_4 + a_4^4,$$
$$B_4 = a_0^3 a_1 + a_0^3 a_3 - a_0 a_3^3 + a_1^3 a_4 + a_1 a_4^3 + a_3 a_4^3 + a_0 a_1^2 a_3 - a_0^2 a_1 a_2 - a_0^2 a_2 a_3 - a_0 a_1 a_4^2 - a_0^2 a_1 a_4 - a_1 a_2 a_4^2 - a_0^2 a_3 a_4 - a_0 a_3 a_4^2 - a_2 a_3 a_4^2 + a_1 a_3^2 a_4 + 2a_0 a_2 a_3 a_4 + 2a_0 a_1 a_2 a_4.$$

For the system to be stable:

$$|\Delta_1| < 0, \quad |A_1| < |B_1|.$$
$$|\Delta_2| > 0, \quad |A_2| > |B_2|.$$
$$|\Delta_3| < 0, \quad |A_3| < |B_3|.$$
$$|\Delta_4| > 0, \quad |A_4| > |B_4|.$$

E. I. Jury
B. H. Bharucha
Dept. of Electrical Engrg.
University of California
Berkeley, Calif.

Information on Translations of Russian Technical Journals

DURING a discussion of the IFAC Moscow Congress at the JACC meeting in Boston, it was mentioned that the Russians knew more about control work in the United States than Americans know about similar work in the USSR because they have excellent translation facilities which are not available to engineers in America. However, Nathan Cohn, representing the Instrument Society of America (ISA) announced that translations of four leading Russian journals have been made by the ISA for the past three years to benefit any individual or organization interested in Russian developments. Yet in spite of the relatively modest cost and the apparent desire to obtain translations of Russian papers, there have been very few subscribers. It was thought that perhaps there has not been enough publicity about this service. Because the Russians have been leaders in the development of control theory, it is important for control engineers to be aware of the progress that they are making. For those who are interested in obtaining more information about these ISA translations, the following ISA announcement is quoted:

" . . . Publication of four Russian technical journals, translated into English, will be continued with the 1960 issues during the coming year by the Instrument Society of America, under a grant from the National Science Foundation. Undertaken as a service to American science and industry, the ISA 'Soviet Instrumentation and Control Translation Series' is now in its fourth year. It affords U. S. scientists and engineers an excellent means to be better informed on the latest developments in the field of Soviet instrumentation. Included in the series are:

"*Automation and Remote Control* (*Avtomatika i Telemekhanika*), considered to be the leading Soviet journal in the automatic control field. Published monthly by the Institute of Automation and Remote Control of the Academy of Science, USSR, it carries approximately 150 pages per issue of articles on all phases of automatic control theories and techniques.

"*Measurement Techniques* (*Izmeritel'naia Tekhnika*), approximately 100 pages per issue, published monthly by the Committee of Standards, Measures and Measuring Instruments of the Council of Ministers, USSR. Of particular interest to those engaged in the study and application of fundamental measurement.

"*Instruments and Experimental Techniques* (*Pribory i Tekhnika Eksperimenta*). Published bimonthly, more than 175 pages per issue, by the Academy of Sciences, USSR, each issue contains articles relating to the function, construction, application and operation of instruments in various fields of instrumentation.

"*Industrial Laboratory* (*Zavodskaya Laboratoriya*). Published monthly by the Ministry of Light Metals, USSR, it contains approximately 125 pages per issue. Articles appearing are on instrumentation for analytical chemistry, and physical and mechanical methods of material research and testing.

"The 1960 translations, as well as translations of previous years published under ISA's program, are available at low subscription rates ranging from $20 to $35 per annual subscription. On a combined order for all four journals special rates apply. Libraries of nonprofit academic institutions are also offered subscriptions at special rates.

"For subscriptions or information write to Foreign Translations Department, Instrument Society of America, 313 Sixth Avenue, Pittsburgh 22, Pennsylvania."

Announcements

Since there were no 1960 WESCON sessions on Automatic Control and no part of the CONVENTION RECORD devoted to Automatic Control, we will not publish any of the WESCON papers in the photo-offset form. Originally, our Group decided to provide a service to its membership by supplying, free of charge, the part of the RECORD specified as Automatic Control. We considered that this was relatively inexpensive for the service it provided, and comments were received from readers expressing appreciation for this service. However, there have been others who object to the photo-offset from in contrast with our regular typeset format. Therefore, since none of the papers were designated as Automatic Control papers *per se* at WESCON, and since they do not appear as a group of papers in the CONVENTION RECORD which could be purchased by any of our members for their control library, we will not publish any of the WESCON papers in typeset form.

PRELIMINARY ANNOUNCEMENT
INTERNATIONAL SYMPOSIUM ON THE TRANSMISSION AND PROCESSING OF INFORMATION

The Professional Group on Information Theory of the Institute of Radio Engineers, in cooperation with the Center of Communication Sciences, Research Laboratory of Electronics, Massachusetts Institute of Technology, is planning to hold an International Symposium on the Transmission and Processing of Information on September 6–8, 1961. This Symposium will be held at the Massachusetts Institute of Technology, Cambridge, Mass.

The purpose of the Symposium will be to provide an outstanding occasion for the presentation of significant new research contributions, of either a theoretical or an experimental nature. As in the case of the similar 1954 and 1956 symposia, no tutorial papers will appear; the program will be planned specifically for active specialists in the field. In order to provide opportunity for creative and thorough discussion, the Symposium *Transactions* will be distributed at least two weeks prior to the meetings.

Submission of papers is hereby invited. In order to carry out the publication plan successfully, the following deadline schedule is necessary:

Receipt of 500–1000-word Abstracts: Immediately.

Receipt of full-length Papers: April 1, 1961.

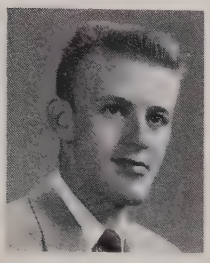
Authors will be notified of the preliminary acceptance of their Abstracts as soon as possible. The final program selection will be made on the basis of the complete Papers, and authors notified by May 1. Abstracts and Papers should be submitted to the Chairman of the Organizing Committee, R. M. Fano, R.L.E., M.I.T., Cambridge 39, Mass.

Additional information about the Symposium will be disseminated as plans develop.

EDWARD M. HOFSTETTER
Asst. Prof. of Elec. Engrg.
Res. Lab. of Electronics
Mass. Inst. Tech.
Cambridge, Mass.

Contributors

D. S. Billingsley was born on June 26, 1929, in Chicago, Ill. He received the B.S. degrees in chemical, industrial, and mechanical engineering in 1950, 1956, and 1956, respectively, and the M.S. degree in chemical engineering in January, 1958, from the Agricultural and Mechanical College of Texas, College Station.



D. S. Billingsley

He served with the U. S. Army from 1951 to 1953, and is presently a graduate student at the A and M College of Texas.

Mr. Billingsley is a member of Phi Lambda Upsilon and an associate member of Sigma Xi.

❖

Kan Chen (S'52–A'55–SM'60) was born on August 28, 1928, in Hongkong, China. He received the B.E.E. degree from Cornell University, Ithaca, N. Y., in 1950 and the M.S. and Sc.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1951 and 1954, respectively.

K. Chen

He joined Westinghouse Electric Corporation, Pittsburgh, Pa., in 1954, and has worked in the fields of solid-state devices, computers and control systems engineering. He is presently Advisory Engineer in the Westinghouse Research Laboratories, Pittsburgh, in charge of a group working on advanced systems techniques. He is also Adjunct Associate Professor in the Electrical Engineering Department of the University of Pittsburgh.

Dr. Chen is a member of Eta Kappa Nu, Tau Beta Pi, Phi Kappa Phi, and Sigma Xi.

❖

Charles A. Desoer (S'50–A'53–SM'57) was born in Brussels, Belgium, in 1926. He received the radio engineer's degree from the University of Liege, Belgium, in 1949, and the Sc.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1953.

C. A. Desoer

He then joined the Bell Telephone Laboratories where he worked in the transmission systems group, principally in the field of network theory. In 1958 he joined the Department of Electrical Engineering at the University of California, Berkeley. His main interest lies in systems and communication theory.

❖

Clarence C. Glover (M'57) was born in Mokane, Mo. on March 14, 1923. He received the B.S.E.E. degree, from the University of Missouri, Columbia, in 1948 and the M.S.E.E. degree from the University of Pittsburgh, Pittsburgh, Pa., in 1952.

C. C. Glover

On obtaining his B.S. degree, Mr. Glover joined the graduate student program of the Westinghouse Electric Corporation. From 1948 to 1952 he worked at the Special Products Division in Pittsburgh, in the field of airborne analog computers, and since 1952 he has been employed at the Air Arm Division in Baltimore, Md. He presently holds the position of Fellow Engineer in the Avionics System Section. He also served as an instructor and did graduate study in the Department of Electrical Engineering at The Johns Hopkins University, Baltimore, during the 1958–1959 and 1959–1960 school years, while on leave of absence from the Westinghouse Electric Corporation.

Mr. Glover is a member of Tau Beta Pi, Eta Kappa Nu, Pi Mu Epsilon, and Sigma Xi.

❖

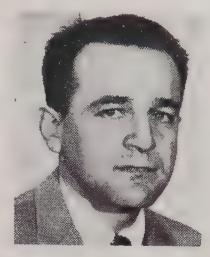
James C. Hung (S'57–M'60) was born in Fukin, China, on February 18, 1929. He received the B.S.E.E. degree from National Taiwan University, Taipei, Taiwan, in 1953, and the M.E.E. degree from New York University, New York, in 1956. At the present time, he is studying toward his doctorate at New York University.

J. C. Hung

From 1954 to 1956, he was a teaching assistant in the Electrical Engineering Department of New York University. From 1956 to the present, he has been an instructor in that department. Besides teaching, he is also a member of the New York University Automatic Control Research Group.

Mr. Hung is a member of Tau Beta Pi and Eta Kappa Nu, and an associate member of the AIEE.

Michael F. Marx was born on June 24, 1922, in Johnsonville, N. Y. He received the B.A.E. and M.A.E. degrees from Rennselaer Polytechnic Institute, Troy, N. Y., in 1944 and 1948, respectively, and the A.E. degree from the California Institute of Technology, Pasadena, in 1949. He received additional training at the General Electric Company, Schenectady, N. Y., where he took courses in servomechanisms, advanced electronic circuits, and transistor circuits.



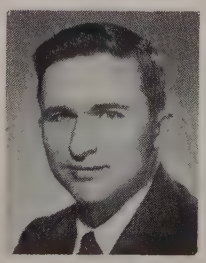
M. F. Marx

From 1943 to 1946, he served with the U. S. Navy as a Fighter Director and CIC Officer. He instructed at the Rensselaer Polytechnic Institute in aerodynamics and aircraft structures from 1946 to 1948. Then, until 1953, he was a senior aerophysics engineer with responsibility for the preliminary stability and control work on the XB-58 and the stability and control of the YB-60 at Convair. Since 1953, he has been a flight control development engineer with the General Electric Company, Schenectady, N. Y. At General Electric, he has been responsible for the bomber wiring layout and the fighter system gradient and computer-flight test correlation of the MX-1137 high-performance flight control. He has been concerned with the Company's self-adaptive flight development and test program, VTOL effort, control techniques and systems for space vehicles, and work on structural feedback. Presently, he is heading the advance engineering flight control development.

Mr. Marx is an Associate Fellow of IAS, and an associate of Sigma Xi, and is a licensed Professional Engineer in New York.

❖

Richard J. McGrath (S'57–M'60) was born in Milwaukee, Wis., on March 25, 1933. He received the B.S.E.E. degree in 1958, the M.S.E.E. degree in 1959, and the Ph.D. degree in 1960, all from the University of Wisconsin, Madison.



R. J. McGrath

He served with the U. S. Army from 1951 through 1953 and has worked summers with General Electric and the Wisconsin Telephone Company.

From June, 1960 to January, 1961, he served as an assistant professor at the University of Wisconsin, mainly carrying on research in the area of adaptive systems. He has since joined the Aerospace Corporation,

in Los Angeles, Calif. His interests lie in the applications of information theory, and computing with emphasis on control and special problems in data handling.

Dr. McGrath is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

❖

Michael G. Rekoff, Jr. (A'56) was born on July 27, 1929, in Galveston, Tex. He received the B.S.E.E. and M.S.E.E. degrees from the Agricultural and Mechanical College of Texas, College Station, in 1951 and 1955, respectively; and the Ph.D. degree, also in electrical engineering, from the University of Wisconsin, Madison, in January of this year.

M. G. Rekoff

From 1951 to 1953, he served in the U. S. Army Signal Corps. He then became an instructor and graduate student at the University of Wisconsin. Presently, he is assistant professor of electrical engineering at the A and M College of Texas.

Dr. Rekoff is a registered Professional Engineer in Wisconsin and Texas and a member of the AIEE and Eta Kappa Nu.

❖

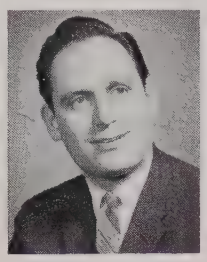
Vincent C. Rideout (M'44–SM'53–F'60) was born in Alberta, Canada, on May 22, 1914. He received the B.Sc. degree in engineering physics from the University of Alberta, Edmonton, in 1938, and the M.S. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1939.

V. C. Rideout

From 1939 to 1946, he was a member of the Bell Telephone Laboratories Technical Staff, working on radar and microwave relay research. In 1946, he joined the staff of the Electrical Engineering Department at the University of Wisconsin, Madison, as an assistant professor. His main interests, both in research and graduate teaching, are now in the fields of computing, network theory and random process studies. During 1954 and 1955, he was in Bangalore, India, where he served as a Technical Cooperation Mission Visiting Professor of Electrical Communication Engineering at the Indian Institute of Science. During 1960 and 1961, he has held a half-time appointment in the Mathematics Research Center of the U. S. Army. He has served as consultant to a number of companies. He now holds the rank of full professor at the University of Wisconsin.

Prof. Rideout is a member of Sigma Xi, Eta Kappa Nu, and Tau Beta Pi.

William C. Schultz (M'59) was born on July 30, 1927, in Sheboygan, Wis. He received the B.S. degree in electrical engineering in 1952, the M.S. degree in electrical engineering in 1953, and the Ph.D. degree in 1958, all from the University of Wisconsin, Madison. His thesis topic was the general subject of control system performance measures.



W. C. Schultz

He served in the U. S. Navy from 1945 to 1948, and taught in the Navy Electronics Technician Training Program from 1946 to 1948. From 1953 to 1955, he served on the staff of the Analog Computing Laboratory at the Allis-Chalmers Manufacturing Company, Milwaukee, Wis., where he contributed to the analysis and design of electrical and hydraulic automatic control systems. From 1955 to 1958, while working toward the Ph.D. degree at the University of Wisconsin, he served as an instructor in electrical engineering; he continued to serve in that capacity until June, 1958, when he was appointed assistant professor. Since July, 1958, he has been with the Cornell Aeronautical Laboratory, Buffalo, N. Y., where he is a principal electronics engineer. His responsibilities have included those of project engineer for a Laboratory project concerned with the Naval Tactical Data System, which included a wide variety of design problems. He has also worked on projects on adaptive flight-control systems, missile trajectory data processing, and control system performance measures.

For the academic year 1960–1961, he is serving as a visiting professor in the School of Electrical Engineering, Cornell University, Ithaca, N. Y. He is teaching courses in automatic control theory, and is active in research and graduate seminar activities.

Dr. Schultz is a member of Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.

❖

Fred B. Smith, Jr., was born in Oak Park, Ill. on January 25, 1930. He received the B.A. degree from Kalamazoo College, Kalamazoo, Mich., in 1952, and the M.S. degree from the University of Illinois, Urbana, in 1954.

F. B. Smith, Jr.

At the University of Illinois, he was a research assistant doing original investigations on nuclear decay schemes. During 1955, he was employed by the Bell Telephone Laboratories working on magnetic core logic studies. During 1956 and 1957, while with the U. S. Army, he was at the Ballistic Research Laboratories, Aberdeen, Md., where he worked on shock tube investigations of weak shock formation and effects. Mr. Smith joined Minneapolis-Honeywell Regulator Co., Minneapolis, Minn., in 1957, and contributed to their first operational digital computer. For the past two years he has been actively working in the areas of adaptive and optimal control systems.

❖

Gerald Weiss (SM'59) was born on August 3, 1922 in Cologne, Germany, and has resided in the United States since 1939. He received the B.E.E. degree from Cooper Union, New York, N. Y., in 1943; the M.S. degree from Harvard University, Cambridge, Mass., in 1948; and the D.E.E. degree from the Polytechnic Institute of Brooklyn, Brooklyn, N. Y., in 1959.

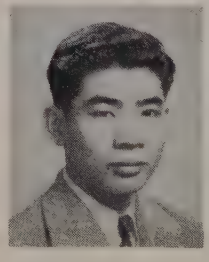
G. Weiss

In 1943, he was employed by the Liquidometer Corporation, Long Island City, N. Y., on the development of electronic fuel gauges. From 1945 to 1947, he served briefly in the U. S. Army and after his discharge he joined the Arma Corporation, Brooklyn, N. Y. At Arma, he was engaged in the development of electromechanical computer components, servomotors, tachometers, magnetic amplifiers, and instrument servomechanisms. In 1953, he joined the W. L. Maxson Corporation, New York, N. Y. as a research engineer, participating in the design of navigation, bombing, and flight path computers; he was subsequently appointed project manager for the ultrasonic flowmeter development program. Since 1958, he has been associated with the Polytechnic Institute of Brooklyn, where he is now Assistant Professor of electrical engineering.

Dr. Weiss is a member of the AIEE, the AACC, Tau Beta Pi and Sigma Xi.

❖

Jack Wing (S'51–A'52–M'57) was born in Sacramento, Calif., on November 28, 1928. He received the B.S.E.E. and M.S.E.E. degrees in 1951 and 1960, respectively, from the University of California, Berkeley, where he is presently studying toward the Ph.D. degree in electrical engineering.



J. Wing

From 1951 to 1952, he worked at Boeing Aircraft Co., Seattle, Wash., as a member of the B-52 electrical system group. Later, from 1953 to 1957, he was employed by the Bendix Radio Division, Towson, Ind., working on surveillance and precision approach radar systems. During 1958, he worked at Lockheed Missile Systems Division, Sunnyvale, Calif., where he was concerned with radar reflectivity studies.

Mr. Wing is a member of Tau Beta Pi and Eta Kappa Nu.